

La comparaison et la simplification de fonctions discriminantes linéaires

THÈSE

PRÉSENTÉE A LA FACULTÉ DES SCIENCES DE L'UNIVERSITÉ DE GENÈVE
POUR OBTENIR LE GRADE DE DOCTEUR EN STATISTIQUE MATHÉMATIQUE

par

Richard-André LANG

de Oftringen (Argovie)

Thèse n° 1311

IMPRIMERIE KUNDIG

GENÈVE

1961

La comparaison et la simplification de fonctions discriminantes linéaires



THÈSE

PRÉSENTÉE A LA FACULTÉ DES SCIENCES DE L'UNIVERSITÉ DE GENÈVE
POUR OBTENIR LE GRADE DE DOCTEUR EN STATISTIQUE MATHÉMATIQUE

par

Richard-André LANG

de Oftringen (Argovie)

Thèse n° 1311

IMPRIMERIE KUNDIG
GENÈVE
1961

La Faculté des Sciences, sur le préavis de M. le professeur
A. LINDER, autorise l'impression de la présente thèse, sans expri-
mer d'opinion sur les propositions qui y sont énoncées.

Genève, le 24 novembre 1959.

Le doyen :

Thèse n° 1311.

Fernand CHODAT.

“For Statistics is pre-eminently a subject, in which the clear realisation of concrete requirements is needed to direct theoretical attention to the essential problems, and to prevent the futile und fruitless elaboration of minor issues. All the great advances to which we can look back, point to this same moral. . . »

R. A. FISHER (1953).

TABLE DES MATIÈRES

	Pages
0 AVANT-PROPOS	325
1 THÉORIE	327
1.1 Introduction	327
1.11 Délimitation du problème. Hypothèses fondamentales	327
1.12 Analyse de la dispersion. Régression et corrélation	328
1.13 La distance généralisée D^2 et le rapport généralisé T	331
1.14 Note historique sur les définitions de D^2 et T	332
1.2 Détermination de la fonction discriminante	334
1.21 Méthode de l'extremum	334
1.22 Méthode de la régression multiple	336
1.23 Analyse de la variance associée à la régression	337
1.3 La distance entre les échantillons	338
1.31 Test de la distance entre deux échantillons	338
1.32 Cas particulier de trois échantillons	340
1.33 La classification d'une valeur individuelle	341
1.4 Comparaisons de fonctions discriminantes	341
1.41 Précision d'une fonction discriminante	341
1.42 Comparaison avec une fonction empirique	347
1.43 Comparaison de deux discriminations calculées	349
1.5 Simplifications de fonctions discriminantes	350
1.51 L'inclusion de nouvelles variables	350
1.52 Importance, pour la discrimination, d'une variable déterminée	351
1.53 Le test de l'écart entre deux distances	352
1.54 La réduction de la distance par élimination d'une variable	352
1.6 Les techniques arithmétiques	355
1.61 Résolution simple par éliminations successives	356
1.62 Détermination simultanée de la fonction discriminante et de la distance associée	358
1.63 L'inversion d'une matrice	361
2 APPLICATION	364
2.1 Les données numériques	364
2.11 Délimitation de la liste primitive	364
2.12 Calculs préliminaires	365
2.13 Problèmes anthropologiques et problèmes statistiques	367
2.2 Les discriminations fondamentales	369
2.21 Les discriminations typiques	369
2.22 Les discriminations raciales complètes	370
2.3 Comparaisons	372
2.31 Comparaison des deux discriminations raciales complètes	372
2.32 Comparaisons basées sur les discriminations typiques	375

	Pages
2.4	Simplifications 376
2.41	Les caractères importants pour la discrimination raciale 376
2.42	Fonctions discriminantes simplifiées 379
2.43	Le rôle des corrélations simples entre caractères 381
2.44	Remarques sur une représentation géométrique 385
2.5	Les valeurs individuelles 385
2.51	Le cas particulier du type II 385
2.52	Classification individuelle 388
2.53	Tests directs utilisant les valeurs individuelles 389
3	RÉCAPITULATION 395
3.11	Résumé 395
3.12	Summary 396
3.13	Zusammenfassung 397
3.2	Liste des symboles et abréviations 398
3.3	Remerciements 402
3.4	Bibliographie 402
3.5	Annexe 405

Adresse permanente de l'auteur:
 RICHARD LANG, Mittelstrasse 14, Langenthal (BE), Suisse.

0 AVANT-PROPOS

Le présent travail doit son origine à la situation bien connue où des investigateurs, en l'occurrence des anthropologistes, ont pris l'initiative d'inviter les statisticiens à collaborer avec eux pour l'élaboration des observations recueillies au cours de l'une de leurs enquêtes. Certaines questions statistiques de base avaient été tirées au clair lors de l'organisation de l'enquête déjà, il est vrai, de même que certaines considérations théoriques — d'un aspect peut-être trop spécifiquement statistique aux yeux des anthropologistes — ont été introduites par les statisticiens au cours de l'élaboration à proprement parler. Cela néanmoins n'a pas empêché l'application statistique, en tant qu'auxiliaire essentiellement pratique, de constituer en tout temps le centre de cette collaboration.

Dans la présente thèse, nous n'avons pas jugé inutile pour autant de faire précéder l'exposé de la méthodologie pratique statistique d'une introduction délibérément théorique dans laquelle nous nous proposons de passer en revue les principaux tests que nous avons été amené à faire intervenir au cours de l'élaboration numérique. C'est donc intentionnellement que nous avons partagé notre travail en deux parties nettement distinctes : une première partie théorique dans laquelle nous avons reconstruit pour les tests spécifiques de l'analyse multidimensionnelle le passage de l'hypothèse à la formule finale prête à l'application pratique, et ceci avec le détail par lequel nous espérons faciliter la compréhension des enchaînements logiques — détail que l'on regrette parfois de ne pas rencontrer dans les travaux originaux. Les méthodes de calcul les plus caractéristiques de l'analyse à plusieurs dimensions y sont également considérées sous leur aspect essentiellement théorique.

La seconde partie de cette thèse sera par contre entièrement réservée à l'exploration numérique d'un matériel anthropométrique par les techniques statistiques précédemment mises au point. Contrairement à la première partie — où par souci de généralité nous utilisons une terminologie mathématique abstraite — nous nous sommes efforcé de rédiger cette seconde partie dans le langage courant des anthropologistes qui auront ainsi l'avantage de voir les conclusions aussi énoncées en termes anthropologiques.

D'emblée nous devons cependant avertir le lecteur, et particulièrement le lecteur anthropologiste, du risque qu'il court en interprétant nos conclusions au-delà de la réponse précise donnée par le test statistique. Rappelons simplement que dans ces recherches anthropologiques, le statisticien avait essentiellement pour tâche d'examiner jusqu'à quel point les données observées confirmaient ou contredisaient certaines hypothèses anthropologiques que, en tant que telles, il ne lui appartient pas de discuter. Le résultat une fois établi, il n'appartenait pas davantage au statisticien de lui attacher une signification autre que celle pouvant être énoncée en termes probabilistes simples.

Note concernant les références. — Pour donner la possibilité de retourner aux travaux originaux et indiquer par là même nos sources, nous n'avons pas hésité d'indiquer régulièrement dans le texte toutes les références bibliographiques. A titre d'information, nous avons énuméré en fin de la liste bibliographique (§ 3.4) quelques publications non mentionnées dans le texte et consacrées à des applications numériques semblables aux nôtres (mensurations sur l'homme vivant); cette énumération ne prétend évidemment pas être exhaustive.

Dans le but de faciliter l'inévitable et fréquent rappel de formules, celles-ci ont été systématiquement munies d'un numéro d'ordre placé entre parenthèses en tête de ligne; la référence se fera selon l'abréviation « eq. » (= équation) ou « form. » (= formule), etc. L'indication du paragraphe ne sera faite, en règle générale, que pour les figures et les

tableaux, notamment ceux qui, parce qu'ils ne sont pas indispensables pour la compréhension du texte, ont été placés en fin du travail (§ 3.5).

Une liste complète de tous les symboles mathématiques, abréviations, conventions d'écriture, etc., avec mention du paragraphe où telle notation sinon sa définition apparaît la première fois, a été dressée; elle constitue le paragraphe 3.2.

1. THÉORIE

1.1 INTRODUCTION

1.11 Délimitation du problème. Hypothèses fondamentales.

L'analyse discriminante, au sens de R. A. FISHER, répond à la question suivante : en présence de plusieurs échantillons de valeurs observées x_{ijk} , quelle est la fonction linéaire :

$$(1) \quad H_{ij} = \sum_{k=1}^p h_k x_{ijk}$$

qui donne le meilleur contraste entre ces échantillons ? Une telle fonction H sera appelée, par définition, fonction discriminante linéaire, ou bien, selon la terminologie de W. L. STEVENS (1945, p. 11), discriminant tout court. Les h_k seront ses coefficients. Nous nous occuperons non seulement du problème de la détermination des coefficients h_k et de leur précision ainsi que du pouvoir séparateur (discriminatoire) de la fonction H , mais encore de la comparaison de différentes fonctions et de la simplification d'un discriminant calculé (cf. § 2.13).

Dans la notation x_{ijk} d'une observation, l'indice i désignera toujours l'échantillon, l'indice j l'individu, et l'indice k (ou l) le nom de la variable (ou la dimension), avec, en général :

$$\begin{aligned} i &= 1, \dots, m \\ j &= 1, \dots, N_i \\ k &= 1, \dots, p, \quad \text{parfois } k = 1, \dots, p + q \\ l &= 1, \dots, p, \quad \text{parfois } l = 1, \dots, p + q. \end{aligned}$$

Cette observation sera d'emblée supposée provenir de l'échantillonnage dans un ensemble de base normal à p dimensions défini par une densité de probabilité φ :

$$(2) \quad \varphi(x_1; \dots; x_k; \dots; x_p) = C \exp(-\Omega/2),$$

Ω étant une forme quadratique :

$$(3) \quad \Omega = \sum_{k=1}^p \sum_{l=1}^p \omega^{kl} (x_k - \mu_k) (x_l - \mu_l)$$

dans laquelle les paramètres μ_k sont les espérances mathématiques :

$$(4) \quad \mu_k = E(x_k).$$

Le coefficient ω^{kl} est l'élément général de l'inverse de la matrice de dispersion (matrice des variances et covariances) $[\omega_{kl}]$ de l'ensemble de base; la constante C est déterminée par l'intégrale :

$$(5) \quad C^{-1} = \int \exp(-\Omega/2) dx_1 dx_2 \dots dx_p,$$

intégrale p -ième étendue à l'espace (euclidien) à p dimensions tout entier (cf. ANDERSON, 1958, p. 11).

Disposant de différents échantillons de valeurs observées, nous préciserons en temps voulu les hypothèses que nous faisons sur $E(x_k)$. En ce qui concerne la dispersion, nous devrons supposer dans tout ce travail l'égalité des variances et covariances dans les divers sous-ensembles de base (hypothèse de l'équidispersion). Sans cette hypothèse, certains tests que nous nous proposons d'établir ne peuvent, en effet, pas être démontrés. Le plus souvent,

nous aurons deux échantillons d'effectifs N_1 et N_2 respectivement; nous ne ferons pas la supposition $N_1 = N_2$ qui, pourtant, aurait l'avantage de simplifier considérablement certaines formules. N désignera, en général, l'effectif total:

$$(6) \quad N = \sum_{i=1}^m N_i ,$$

tandis que nous réservons la notation n pour les degrés de liberté d'une dispersion (variance, covariance) estimée.

Sauf indication contraire, les indices de sommation parcourent toujours tout l'ensemble des valeurs qu'ils peuvent prendre dans la démonstration en cours; dans toutes les sommations, nous pourrions donc sous-entendre la limite supérieure de l'indice, à moins qu'il y ait une raison déterminée de la préciser. En particulier, les doubles sommes seront souvent écrites sous la forme condensée:

$$\sum_{i,j}, \sum_{k,l} \text{ au lieu de } \sum_{i=1}^m \sum_{j=1}^{N_i}, \sum_{k=1}^p \sum_{l=1}^p \text{ par exemple.}$$

De cette façon, nous pourrions écrire quelques égalités élémentaires sous la forme:

$$(7) \quad x_k = \sum_{i,j} x_{ijk}/N, \quad \text{moyenne générale de la variable } k\text{-ième et pour la réunion de tous les échantillons; } k = 1, \dots, p;$$

$$(8) \quad \bar{x}_{ik} = \sum_j x_{ijk}/N_i, \quad \text{moyenne de la variable } k\text{-ième dans l'échantillon } i\text{-ième; } i = 1, \dots, m; k = 1, \dots, p;$$

$$(9) \quad x_k = \sum_i N_i \bar{x}_{ik}/N, \quad \text{moyenne générale sous forme de moyenne pondérée des moyennes d'échantillon; } k = 1, \dots, p.$$

De même pour les valeurs H_{ij} de la fonction discriminante:

$$(10) \quad \bar{H} = \sum_k h_k \bar{x}_k, \quad \text{moyenne générale;}$$

$$(11) \quad \bar{H}_i = \sum_k h_k \bar{x}_{ik}, \quad \text{moyenne de l'échantillon } i\text{-ième; } i = 1, \dots, m;$$

$$(12) \quad \bar{H} = \sum_i N_i \bar{H}_i/N, \quad \text{moyenne pondérée des moyennes d'échantillon.}$$

Le double indice supérieur indiquera, en règle générale, l'opération d'inversion matricielle, donc par exemple:

$$(13) \quad [w^{kl}] = [w_{kl}]^{-1};$$

les crochets [...] dénoteront systématiquement les matrices.

Toutes les égalités dépendant explicitement soit de l'indice k soit de l'indice l , sont valables, en principe, quelle que soit la valeur de cet indice.

1.12 Analyse de la dispersion. Régression et corrélation.

L'information, concernant la dispersion (variance, covariance) de m échantillons d'effectifs N_i ($i = 1, \dots, m$) de valeurs observées x_{ijk} peut, par analogie avec une analyse de variance à une seule dimension, être subdivisée en deux parties selon le schéma du tableau 1. Suivant la terminologie de P. C. MAHALANOBIS, on parlera alors de l'analyse de la dispersion.

TABLEAU 1

Variabilité	Degrés de liberté	Sommes des carrés ou des produits
entre échantillons	$m - 1$	$S'_{kl} = \sum_{i,j} (\bar{x}_{ik} - \bar{x}_k) (\bar{x}_{il} - \bar{x}_l)$
résiduelle	$N - m$	$S''_{kl} = \sum_{i,j} (x_{ijk} - \bar{x}_{ik}) (x_{ijl} - \bar{x}_{il})$
totale	$N - 1$	$S_{kl} = \sum_{i,j} (x_{ijk} - \bar{x}_k) (x_{ijl} - \bar{x}_l)$

La variabilité résiduelle est aussi appelée variabilité à l'intérieur des échantillons par opposition à la variabilité entre échantillons. En posant :

$$(14) \quad \text{Var} (x_k) = S_{kk}/(N - 1) \quad (\text{variance totale}),$$

$$(15) \quad \text{Cov} (x_k; x_l) = S_{kl}/(N - 1) \quad (\text{covariance totale}),$$

on a pour définition du coefficient de corrélation (simple) totale $r (x_k; x_l)$ entre les variables x_k et x_l :

$$(16) \quad r (x_k; x_l) = \text{Cov} (x_k; x_l) \{ \text{Var} (x_k) \text{Var} (x_l) \}^{-\frac{1}{2}}, \quad \text{ou bien :}$$

$$(17) \quad r^2 = S_{kl}^2 / (S_{kk} S_{ll}).$$

(Cf. par exemple FISHER, 1958, p. 183).

De façon semblable, en posant :

$$(18) \quad \text{var} (x_k) = S''_{kk}/(N - m) \quad (\text{variance à l'intérieur des échantillons})$$

$$(19) \quad \text{cov} (x_k; x_l) = S''_{kl}/(N - m) \quad (\text{covariance à l'intérieur des échantillons}),$$

on aura pour définition du coefficient de corrélation (simple) $r'' (x_k; x_l)$ à l'intérieur des échantillons, entre les variables x_k et x_l :

$$(20) \quad r'' (x_k; x_l) = \text{cov} (x_k; x_l) \{ \text{var} (x_k) \text{var} (x_l) \}^{-\frac{1}{2}}, \quad \text{ou bien :}$$

$$(21) \quad r''^2 = S_{kl}''^2 / (S_{kk}'' S_{ll}'').$$

Pour les valeurs individuelles $H_{ij} = \sum_k h_k x_{ijk}$ d'un discriminant, on peut établir une analyse de variance telle qu'elle se trouve résumée dans le tableau 2. La question des degrés de liberté sera élucidée aux paragraphes 1.23 et 1.31. Les relations de ce tableau 2 s'obtiennent par simples transformations algébriques à partir des définitions.

TABLEAU 2

Variabilité	Degrés de liberté	Somme des carrés
entre échantillons	...	$S'_H = \sum_{i,j} (\bar{H}_i - \bar{H})^2 = \sum_{k,l} h_k h_l S'_{kl}$
résiduelle	...	$S''_H = \sum_{i,j} (H_{ij} - \bar{H}_i)^2 = \sum_{k,l} h_k h_l S''_{kl}$
totale	...	$S_H = \sum_{i,j} (H_{ij} - \bar{H})^2 = \sum_{k,l} h_k h_l S_{kl}$

Soit y_{ij} une nouvelle variable observable sur l'individu $(i; j)$. La régression linéaire multiple Y de y en $x_1, \dots, x_k, \dots, x_p$ sera:

$$(22) \quad Y_{ij} - \bar{y} = \sum_k b_k (x_{ijk} - \bar{x}_k)$$

où b_k est le coefficient de régression multiple défini par les équations:

$$(23) \quad \sum_k b_k S_{kl} = S_{ly}, \quad (l = 1, \dots, p), \text{ ou bien sous forme matricielle:}$$

$$(24) \quad [S_{kl}] [b_k] = [S_{ly}], \text{ avec:}$$

$$(25) \quad \bar{y} = \sum_{i,j} y_{ij} / N,$$

$$(26) \quad S_{ly} = \sum_{i,j} (y_{ij} - \bar{y}) (x_{ijl} - \bar{x}_l).$$

La variabilité totale:

$$(27) \quad S_{yy} = \sum_{i,j} (y_{ij} - \bar{y})^2$$

peut être scindée en deux parties comme il l'est indiqué au tableau 3. La variabilité résiduelle s'appelle aussi variabilité autour de la régression.

TABLEAU 3

Variabilité	Degrés de liberté	Somme des carrés
régression	p	$\sum_{i,j} (Y_{ij} - \bar{y})^2 = \sum_k b_k S_{ky} = r_p^2 S_{yy}$
résiduelle	$N - p - 1$	$\sum_{i,j} (Y_{ij} - y_{ij})^2 = S_{yy} - \sum_k b_k S_{ky} = (1 - r_p^2) S_{yy}$
totale	$N - 1$	$\sum_{i,j} (y_{ij} - \bar{y})^2 = S_{yy}$

Rappelons la définition du coefficient r_p de corrélation multiple (cf. FISHER, 1958, p. 258):

$$(28) \quad r_p^2 = \frac{\sum_{i,j} (Y_{ij} - \bar{y})^2}{\sum_{i,j} (y_{ij} - \bar{y})^2} = \frac{\sum_k b_k S_{ky}}{S_{yy}}.$$

D'autre part, le coefficient de détermination B sera défini par:

$$(29) \quad B = \frac{\sum_{i,j} (Y_{ij} - \bar{y})^2}{\sum_{i,j} (y_{ij} - \bar{y})^2}$$

aussi bien dans le cas d'une régression (linéaire) simple que dans le cas multiple. C'est pourquoi l'on aura:

$$(30) \quad B = r_p^2 \quad \text{dans la régression multiple, mais par contre:}$$

$$(31) \quad B = r^2(x; y) \quad \text{dans le cas simple (cf. par exemple EZEKIEL & FOX, 1959, p. 130).}$$

Ce coefficient de détermination mesure donc la portion de la variabilité totale explicable à travers la régression linéaire, par les seules variations des variables x indépendantes.

Remarquons ici et une fois pour toutes que nous désignerons régulièrement par $F(f_1; f_2)$ le rapport de deux variances estimées. Nous empruntons à HALD la notation $(f_1; f_2)$ pour les degrés de liberté respectifs, mais nous préférons conserver à cette répartition la notation F presque universellement adoptée de nos jours et qui est due à SNEDECOR (cf. HALD, 1952a, p. 375; HALD, 1952b, p. 26; SNEDECOR, 1956, p. 244). Nous utiliserons les abréviations 0 (en exposant), * et ** pour désigner une valeur respectivement non significative, significative (au seuil $P = 0.05$) ou bien hautement significative ($P = 0.01$).

1.13 La distance généralisée D^2 et le rapport généralisé T .

Soit deux échantillons de valeurs observées x_{ijk} ($i = 1$ ou 2 ; $j = 1, \dots, N_i$; $k = 1, \dots, p$). Un fréquent usage sera fait de la distance généralisée D_p^2 de MAHALANOBIS définie ici (cf. § 1.14) par la relation :

$$(32) \quad D_p^2 = \sum_{k,l} w^{kl} (\bar{x}_{1k} - \bar{x}_{2k}) (\bar{x}_{1l} - \bar{x}_{2l})$$

avec $[w^{kl}]$ comme inverse de la matrice de dispersion estimée $[w_{kl}]$, c'est-à-dire :

$$(33) \quad \begin{aligned} nw_{kl} &= \sum_{i,j} (x_{ijk} - \bar{x}_{ik}) (x_{ijl} - \bar{x}_{il}), \\ n &= N_1 + N_2 - 2 \\ k &= 1, \dots, p; l = 1, \dots, p \quad (\text{cf. RAO, 1952, p. 246}). \end{aligned}$$

Telle quelle, la valeur de la distance peut être considérée comme estimation du paramètre :

$$(34) \quad \Delta_p^2 = \sum_{k,l} \omega^{kl} (\mu_{1k} - \mu_{2k}) (\mu_{1l} - \mu_{2l})$$

défini dans l'ensemble de base, avec $[w^{kl}]$ comme inverse de la matrice de dispersion de cet ensemble de base, et :

$$(35) \quad \begin{aligned} \mu_{ik} &= E_{(j)}(x_{ijk}), \\ i &= 1 \text{ ou } 2; k = 1, \dots, p; l = 1, \dots, p \end{aligned}$$

(cf. KENDALL, 1946, p. 359). On peut montrer que la distance D_p^2 est invariante à travers les transformations linéaires des variables x_k (BOSE & ROY, 1938, p. 21).

Si, par ailleurs, $d_1, \dots, d_p$ sont les variables d'une répartition normale à p dimensions ayant une matrice de dispersion $[\alpha_{kl}]$ dont on possède une estimation $[a_{kl}]$ à n degrés de liberté, on définit ici (cf. § 1.14) le rapport généralisé T_p de HOTELLING par la relation :

$$(36) \quad nT_p = \sum_{k,l} a^{kl} d_k d_l.$$

Lorsqu'on fait l'hypothèse :

$$(37) \quad E(d_1) = 0, \dots, E(d_k) = 0, \dots, E(d_p) = 0,$$

on peut montrer que :

$$(38) \quad (n - p + 1) T_p / p = F(p; n - p + 1).$$

Si l'on introduit encore q variables supplémentaires $d_{p+1}, \dots, d_{p+q}$ dont la répartition ne dépend pas de celle des $d_1, \dots, d_p$, et en faisant l'hypothèse :

$$(39) \quad \begin{aligned} E(d_{p+1}) &= 0, \dots, E(d_{p+k'}) = 0, \dots, E(d_{p+q}) = 0, \\ 1 &\leq k' \leq q, \end{aligned}$$

mais sans que nécessairement $E(d_k) = 0$ pour $1 \leq k \leq p$, on peut de même démontrer que :

$$(40) \quad \frac{n - p - q + 1}{q} \left\{ \frac{1 + T_{p+q}}{1 + T_p} - 1 \right\} = F(q; n - p - q + 1),$$

(RAO, 1952, p. 238). Nous utiliserons ces deux résultats sans les démontrer. On peut également établir que le rapport T_p est invariant à travers les transformations linéaires des variables d_k (RAO, 1952, § 2d.2).

Posons :

$$(41) \quad d_k = \bar{x}_{1k} - \bar{x}_{2k} \quad (k = 1, \dots, p).$$

La matrice de dispersion estimée $[a_{kl}]$ des observations d_k ainsi définies sera :

$$(42) \quad [a_{kl}] = (1/\eta) [w_{kl}], \text{ avec :}$$

$$(43) \quad \eta = N_1 N_2 / (N_1 + N_2) \quad \text{pour double moyenne harmonique.}$$

On a l'égalité:

$$(44) \quad nT_p = \sum_{k,l} \eta w^{kl} (\bar{x}_{1k} - \bar{x}_{2k}) (\bar{x}_{1l} - \bar{x}_{2l}) \\ = \eta D_p^2; \quad \text{autrement dit:}$$

$$(45) \quad T_p = \nu D_p^2,$$

$$(46) \quad \nu = N_1 N_2 / \{ (N_1 + N_2) (N_1 + N_2 - 2) \}.$$

Cela établit une relation entre la distance D_p^2 et le rapport T_p , de sorte que tous les tests concernant la quantité T_p pourront être interprétés en termes de distance généralisée D_p^2 .

Il est à noter que dans le cas d'une seule dimension ($p = 1$), le test:

$$(n - p + 1) T_p = F(p; n - p + 1) \quad (\text{cf. eq. 38})$$

est identique au test classique de « STUDENT » pour l'écart entre les moyennes de deux échantillons d'observations x_{ijk} ($i = 1$ ou 2 ; $j = 1, \dots, N_i$; $k = 1$). En effet:

$$(47) \quad a_{11} = w_{11}/\eta \\ = \sum_{i,j} (x_{ij1} - \bar{x}_{11})^2/n,$$

$$(48) \quad T_1 = \eta (\bar{x}_{11} - \bar{x}_{21}) / \{ \sum_{i,j} (x_{ij1} - \bar{x}_{11})^2/n \}.$$
 En posant:

$$(49) \quad s^2 = \sum_{i,j} (x_{ij1} - \bar{x}_{11})^2/n,$$

$$(50) \quad \bar{x}_{11} = \bar{x}, \quad \bar{x}_{21} = \bar{x}', \quad \text{on retrouve:}$$

$$(51) \quad \frac{|\bar{x} - \bar{x}'|}{s} \sqrt{\frac{N_1 N_2}{N_1 + N_2}} = \sqrt{F(1; N_1 + N_2 - 2)} \\ = t_{[N_1 + N_2 - 2]},$$

qui est précisément la forme habituelle du test de l'écart entre deux moyennes, par la répartition t de « STUDENT » (cf. par exemple FISHER, 1958, § 24.1).

1.14 Note historique sur les définitions de D^2 et de T .

Partant d'une première forme provisoire de la distance entre deux échantillons p -dimensionnels d'effectifs N_1 et N_2 , à savoir:

$$(52) \quad pD = \sum_k d_k^2/s_k^2,$$

où d_k est l'écart entre les moyennes du premier et du second échantillon pour la k -ième des p variables, et s_k un écart-type moyen de cette k -ième variable établi sur un grand nombre d'échantillons différents, MAHALANOBIS proposa par la suite un coefficient D' légèrement modifié:

$$(53) \quad D' = D - 1/\eta, \\ 1/\eta = 1/N_1 + 1/N_2,$$

coefficient préférable au précédent pour certaines raisons théoriques (MAHALANOBIS, 1927, Appendix 3). A noter la conservation du facteur p provenant de ce que, historiquement, les recherches de MAHALANOBIS représentent le prolongement logique des investigations de K. PEARSON, créateur du coefficient de ressemblance raciale (*CRL*, *Coefficient of Racial Likeness*) défini comme:

$$(54) \quad CRL = (1/p) \sum_k \{ d_k^2/(\sigma_{1k}^2/N_1 + \sigma_{2k}^2/N_2) \} - 1,$$

abstraction faite d'une correction additive $1/p$ (PEARSON, 1926, p. 107).

Par la suite, la distance généralisée D^2 (entre deux échantillons) devait être définie, à proprement parler, par MAHALANOBIS sous la forme non studentisée:

$$(55) \quad pD^2 = \sum_{k,l} (\beta^{kl} d_k d_l) - 1/\eta, \quad \text{avec:}$$

$$(56) \quad \beta_{kl} = N_2 \sigma_{1k} \sigma_{1l} \rho_{1kl} / (N_1 + N_2) + N_1 \sigma_{2k} \sigma_{2l} \rho_{2kl} / (N_1 + N_2)$$

$$(57) \quad \sigma_{ik} = \left. \begin{array}{l} \text{déviations-standard de la} \\ \text{variable } x_k \end{array} \right\} \quad \text{dans l'ensemble de base } (i = 1 \text{ ou } 2),$$

$$(58) \quad \rho_{ikl} = \left. \begin{array}{l} \text{corrélations des variables} \\ x_k \text{ et } x_l \end{array} \right\}$$

$$(59) \quad [\beta^{kl}] = [\beta_{kl}]^{-1}$$

(cf. BOSE, 1936, p. 146; MAHALANOBIS, 1936, pp. 49-55).

Enfin, la forme studentisée de la distance D^2 (entre deux échantillons) est définie comme:

$$(60) \quad pD^2 = \sum_{k,l} s^{kl} d_k d_l,$$

où s_{kl} est basée sur la réunion des sommes de produits corrigées à l'intérieur des deux échantillons, avec:

$$(61) \quad [s^{kl}] = [s_{kl}]^{-1} \quad \text{comme convenu (cf. BOSE & ROY, 1938, p. 20).}$$

Conformément à RAO cependant, nous avons adopté dans tout le présent travail pour définition de la distance généralisée D_p^2 entre deux échantillons la grandeur:

$$(62) \quad D_p^2 = \sum_{k,l} s^{kl} d_k d_l$$

en laissant donc tomber le facteur p (RAO, 1952, p. 246).

En ce qui concerne la définition du rapport généralisé T^2 , elle s'écrivait à l'origine (HOTELLING, 1931, pp. 364-365) sous la forme:

$$(63) \quad T^2 = \sum_{k,l} w^{kl} \xi_k \xi_l,$$

en notations légèrement modifiées pour éviter les confusions, et l'on avait:

$$(64) \quad nw_{kl} = \sum_{i,j} (x_{ijk} - \bar{x}_{ik})(x_{ijl} - \bar{x}_{il}),$$

$$n = N_1 + N_2 - 2,$$

$$(65) \quad \xi_k = \sqrt{\eta} (\bar{x}_{1k} - \bar{x}_{2k}).$$

En remarquant que la matrice $[a_{kl}]$ de la définition T_p (eq. 36) est liée à la matrice $[w_{kl}]$ ci-dessus par la relation:

$$(66) \quad [a_{kl}] = (1/\eta) [w_{kl}],$$

on est conduit à l'égalité:

$$(67) \quad T^2 = nT_p.$$

Par ailleurs, le test du rapport T^2 s'énonçait, dans les hypothèses faites (cf. HOTELLING, 1931, p. 377) sous la forme:

$$(68) \quad z = \frac{1}{2} L_n (n - p + 1) T^2 - \frac{1}{2} L_n (np),$$

degrés de liberté: p et $n - p + 1$, en d'autres termes:

$$(69) \quad z = L_n \sqrt{(n - p + 1) T^2 / np}; \quad L_n \text{ désigne le logarithme naturel.}$$

Rappelons que la répartition z se ramène à celle de F par la transformation:

$$(70) \quad z = L_n \sqrt{F},$$

de sorte que le test en question s'énonce de la manière suivante:

$$(71) \quad (n - p + 1) T_p/p = F(p; n - p + 1) \quad \text{conformément à l'équation 38.}$$

Les formules du présent travail sont donc parfaitement cohérentes avec les données de la littérature exprimées parfois dans une notation simplement plus ancienne.

1.2 DÉTERMINATION DE LA FONCTION DISCRIMINANTE

1.21 Méthode de l'extremum.

Soit $H = \sum_k h_k x_k$ le discriminant dont les valeurs individuelles seraient données par $H_{ij} = \sum_k h_k x_{ijk}$. Nous déterminons les coefficients h_k de manière que le rapport:

$$(72) \quad \theta(h_1; \dots; h_p) = \frac{\sum_{i,j} (\bar{H}_i - \bar{H})^2}{\sum_{i,j} (H_{ij} - \bar{H}_i)^2} \\ = S'_H / S''_H,$$

considéré comme fonction des coefficients h_k , soit maximum. C'est la méthode bien connue d'incorporer au discriminant la propriété de fluctuer faiblement à l'intérieur des échantillons, mais fortement entre les divers échantillons.

En désignant par ∂_k la dérivation partielle par rapport à la variable h_k , on doit donc satisfaire aux conditions:

$$(73) \quad \partial_k \theta(h_1; \dots; h_p) = 0 \quad \text{pour tout } k = 1, \dots, p.$$

Autrement dit, en supposant $S''_H \neq 0$:

$$(74) \quad S''_H \partial_k S'_H - S'_H \partial_k S''_H = 0. \quad \text{Comme:}$$

$$(75) \quad \partial_k S'_H = 2 \sum_l h_l S'_{kl},$$

$$(76) \quad \partial_k S''_H = 2 \sum_l h_l S''_{kl},$$

on aboutit, en substituant dans les équations 74, aux relations:

$$(77) \quad \sum_l h_l (S'_{kl} - \theta S''_{kl}) = 0 \quad (k = 1, \dots, p), \text{ ou bien, sous forme matricielle:}$$

$$(78) \quad [S'_{kl} - \theta S''_{kl}] [h_l] = [0],$$

avec $[0]$ pour matrice nulle.

Pour que ce système d'équations linéaires homogènes admette une solution $h_1, \dots, h_p$ non identiquement nulle, c'est-à-dire un vecteur-colonne $[h_l]$ non nul, il faut que le déterminant des coefficients soit nul:

$$(79) \quad || S'_{kl} - \theta S''_{kl} || = 0.$$

La plus grande racine θ de cette équation déterminantale introduite dans le système homogène (eq. 77 ou 78), conduira à la solution. Celle-ci, comme dans tout système homogène, ne pourra être fixée qu'à une constante multiplicative près. Cette démonstration est valable pour un nombre arbitraire d'échantillons ($i = 1, 2, 3, \dots, m$). On peut montrer qu'il existe toujours une racine θ positive au moins (LETESTU, 1948, p. 22).

Si nous nous limitons maintenant au cas particulier de $m = 2$ échantillons d'effectifs N_1 et N_2 , la condition d'extremum de la fonction θ revient à exiger que l'écart $\bar{H}_1 - \bar{H}_2$ soit aussi grand que possible par rapport à la variabilité à l'intérieur des échantillons; car en effet, on a la relation:

$$(80) \quad S'_H = \sum_{i,j} (\bar{H}_i - \bar{H})^2 = \gamma (\bar{H}_1 - \bar{H}_2)^2$$

si l'on prend soin de substituer :

$$\bar{H} = (N_1 \bar{H}_1 + N_2 \bar{H}_2)/N \quad \text{selon eq. 12; et:}$$

$$\eta = N_1 N_2 / (N_1 + N_2) \quad \text{selon eq. 43. Posons encore:}$$

$$(81) \quad d_k = \bar{x}_{1k} - \bar{x}_{2k},$$

$$(82) \quad d_H = \bar{H}_1 - \bar{H}_2 = \sum_k h_k d_k. \quad \text{Alors, on obtient:}$$

$$(83) \quad S'_{kl} = \eta d_k d_l,$$

$$(84) \quad S'_H = \eta \left(\sum_k (h_k d_k)^2 \right) = \eta d_H^2, \quad \text{et enfin:}$$

$$(85) \quad \theta = \eta d_H^2 / S''_H.$$

En introduisant ces résultats dans le système homogène (eq. 77) mis sous la forme :

$$(86) \quad \sum_l h_l S'_{kl} = \theta \sum_l h_l S''_{kl}, \quad \text{on trouve:}$$

$$(87) \quad \eta \sum_l h_l d_k d_l = \eta (d_H^2 / S''_H) \sum_l h_l S''_{kl} \quad (k = 1, \dots, p), \quad \text{ou bien:}$$

$$(88) \quad \sum_l h_l S''_{kl} = d_k (S''_H / d_H) \quad (k = 1, \dots, p).$$

Comme, par la nature même du problème, les inconnues $h_1, \dots, h_p$ ne sont déterminées qu'à une constante multiplicative près, on peut négliger le facteur S''_H / d_H au second membre en supposant toujours $d_H \neq 0$. On a alors simplement à résoudre :

$$(89) \quad \sum_l h_l S''_{kl} = d_k \quad (k = 1, \dots, p), \quad \text{c'est-à-dire:}$$

$$(90) \quad [S''_{kl}] [h_l] = [d_k] \quad \text{sous forme matricielle, ou encore:}$$

$$(91) \quad \sum_l h_l w_{kl} = d_k \quad (k = 1, \dots, p), \quad \text{c'est-à-dire:}$$

$$(92) \quad [w_{kl}] [h_l] = [d_k] \quad \text{sous forme matricielle, avec:}$$

$$(93) \quad w_{kl} = S''_{kl} / (N - 2).$$

On peut en effet remarquer que seul le rapport h_l / h_k ($k \neq l$) est univoquement déterminable. En d'autres termes, tout système $[h'_k]$ proportionnel à une solution $[h_k]$ convient également en tant que solution au problème de la discrimination. Car si λ est une constante de proportionnalité, et que :

$$(94) \quad h'_k = \lambda h_k \quad (k = 1, \dots, p), \quad \text{on peut vérifier directement que:}$$

$$(95) \quad \theta(h'_1; \dots; h'_p) \equiv \theta(h_1; \dots; h_p) \quad \text{quelle que soit } \lambda \neq 0.$$

On peut encore dire qu'en multipliant tous les coefficients d'un discriminant par un facteur constant non nul, on obtient un nouveau discriminant équivalent au premier.

Observons encore qu'au lieu de discuter l'équation 79 pour θ , on peut poser :

$$(96) \quad S''_{kl} = S_{kl} - S'_{kl} \quad (\text{cf. tabl. 1, § 1.21), et se ramener ainsi à la forme:}$$

$$(97) \quad (1 + \theta)^p \parallel S'_{kl} - S_{kl} \theta / (1 + \theta) \parallel = 0.$$

Du moment que la fonction θ n'est jamais négative en tant que rapport de deux sommes de carrés, $(1 + \theta)^p$ est toujours positif, de sorte qu'il suffira de résoudre :

$$(98) \quad \parallel S'_{kl} - \theta' S_{kl} \parallel = 0, \quad \text{avec:}$$

$$(99) \quad \theta' = \theta / (1 + \theta), \quad 0 \leq \theta' \leq 1.$$

Pour le calcul pratique, cette dernière inégalité prend une certaine importance. Une fois la plus grande racine θ' déterminée, il suffira de la substituer dans le système :

$$(100) \quad [S'_{kl} - \theta' S_{kl}] [h_l] = [0]$$

pour obtenir la solution $[h_k]$ comme précédemment. (Cf. STEVENS, 1945, p. 22).

En ce qui concerne l'unicité de la solution des équations 89-92, l'on se référera aux travaux de BARTLETT (par exemple BARTLETT, 1938, p. 37).

1.22 Méthode de la régression multiple.

Considérons deux échantillons d'effectifs N_1 et N_2 , ($N_1 + N_2 = N$), et introduisons une pseudo-variable y_{ij} en définissant:

$$(101) \quad y_{1j} = N_2/N \text{ pour le premier échantillon } (j = 1, \dots, N_1),$$

$$y_{2j} = -N_1/N \text{ pour le second échantillon } (j = 1, \dots, N_2), \text{ de sorte que:}$$

$$(102) \quad \bar{y} = \{N_1 N_2/N + N_2 (-N_1)/N\}/N = 0.$$

On peut alors déterminer la régression linéaire multiple de y_{ij} en les variables x_{ijk} , c'est-à-dire:

$$(103) \quad Y_{ij} = \sum_k b_k (x_{ijk} - \bar{x}_k)$$

où les coefficients b_k satisfont aux équations suivantes:

$$(104) \quad \sum_k b_k S_{kl} = S_{ly} \quad (l = 1, \dots, p). \text{ Or:}$$

$$(105) \quad S_{ly} = \sum_{j=1}^{N_1} y_{1j} (x_{1jl} - \bar{x}_l) + \sum_{j=1}^{N_2} y_{2j} (x_{2jl} - \bar{x}_l) \\ = \{N_2 \sum_{j=1}^{N_1} (x_{1jl} - \bar{x}_l) - N_1 \sum_{j=1}^{N_2} (x_{2jl} - \bar{x}_l)\}/N.$$

En introduisant maintenant la relation:

$$(106) \quad \bar{x}_l = (N_1 \bar{x}_{1l} + N_2 \bar{x}_{2l})/N, \text{ et en posant de nouveau:}$$

$$\eta = N_1 N_2/N,$$

$$d_l = \bar{x}_{1l} - \bar{x}_{2l} \quad (l = 1, \dots, p), \text{ on aboutit en fin de compte à:}$$

$$(107) \quad S_{ly} = \eta (\bar{x}_{1l} - \bar{x}_{2l}) = \eta d_l, \text{ de sorte que le système d'équations pour les coefficients } b_k \text{ (eq. 104) prend la forme suivante:}$$

$$(108) \quad \sum_k b_k S_{kl} = \eta d_l \quad (l = 1, \dots, p). \text{ Observons que:}$$

$$S_{kl} = S'_{kl} + S''_{kl} \quad (\text{cf. tabl. 1, § 1.12}), \text{ et que:}$$

$$S'_{kl} = \eta d_k d_l \quad (\text{eq. 83}), \text{ de sorte que:}$$

$$(109) \quad \sum_k b_k S''_{kl} = d_l \{ \eta (1 - \sum_k b_k d_k) \} \quad (l = 1, \dots, p).$$

Le facteur:

$$(110) \quad \varepsilon = \eta (1 - \sum_k b_k d_k) \text{ ne dépend plus des indices } k \text{ et } l.$$

En rapprochant le système d'équations 109 au système 89 précédemment établi, on constate l'équivalence des deux systèmes à une constante multiplicative ε près, mais qui est sans importance pour la discrimination (cf. § 1.21):

$$(111) \quad \sum_k b_k S''_{kl} = \varepsilon d_l \quad \left\{ \right. \\ (112) \quad \sum_k b_k S''_{kl} = d_l \quad \left. \right\} \quad (l = 1, \dots, p).$$

Autrement dit, les coefficients b_k de cette régression particulière peuvent être pris pour coefficients d'un discriminant (cf. KENDALL, 1946, p. 344).

La relation exacte entre les coefficients b_k et h_k est la suivante :

$$(113) \quad h_k = b_k / \varepsilon .$$

En vertu de l'équation 119 (§ 1.23), et de la définition du coefficient de corrélation multiple r_p^2 (§ 1.12), on peut écrire en fin de compte :

$$(114) \quad r_p^2 = \sum_k b_k d_k , \quad \text{de sorte que :}$$

$$(115) \quad \varepsilon = \eta (1 - r_p^2) , \quad \text{et puis :}$$

$$(116) \quad b_k = \eta (1 - r_p^2) h_k \quad (\text{cf. FISHER, 1938, p. 377}).$$

1.23 Analyse de la variance associée à la régression.

Du moment que nous avons calculé une régression, on pourrait être tenté de lui associer une analyse de variance équivalant, formellement, à la décomposition de la variabilité totale $\sum_{i,j} (y_{ij} - \bar{y})^2$ en une partie attribuable aux variations des variables indépendantes x_k , c'est-à-dire $\sum_{i,j} (Y_{ij} - \bar{y})^2$, et une variabilité résiduelle $\sum_{i,j} (y_{ij} - Y_{ij})^2$ (cf. FISHER, 1938, p. 377). On aurait ainsi :

$$(117) \quad \begin{aligned} \sum_{i,j} (y_{ij} - \bar{y})^2 &= \sum_{i,j} y_{ij}^2 = \sum_{j=1}^{N_1} y_{1j}^2 + \sum_{j=1}^{N_2} y_{2j}^2 \\ &= N_1 (N_2/N)^2 + N_2 (N_1/N)^2 \\ &= (N_1 + N_2) N_1 N_2 / N^2 = \eta , \end{aligned}$$

en tenant compte de la définition de y_{ij} (eq. 101), de celle de η (eq. 43) et du fait que $\bar{y}$ est nulle. On a d'autre part :

$$(118) \quad \begin{aligned} \sum_{i,j} (Y_{ij} - \bar{y})^2 &= \sum_{i,j} Y_{ij}^2 = \sum_{i,j} \left\{ \sum_k b_k (x_{ijk} - \bar{x}_k) \right\}^2 \\ &= \sum_{i,j} \sum_k b_k \sum_l b_l (x_{ijk} - \bar{x}_k) (x_{ijl} - \bar{x}_l) \\ &= \sum_k b_k \sum_l b_l S_{kl} \\ &= \sum_k b_k S_{ky} \\ &= \eta \sum_k b_k d_k , \end{aligned}$$

en utilisant les équations 108 qui déterminent les coefficients b_k , et les relations (eq. 107) qui découlent, pour S_{ky} , de la définition particulière des y_{ij} . Des relations 117 et 118, on déduit :

$$(119) \quad \begin{aligned} r_p^2 &= \sum_{i,j} (Y_{ij} - \bar{y})^2 / \sum_{i,j} (y_{ij} - \bar{y})^2 \\ &= \eta \sum_k b_k d_k / \eta \\ &= \sum_k b_k d_k . \end{aligned}$$

On peut donc construire l'analyse de variance conformément au tableau 4. On notera l'étroite correspondance avec le tableau 3.

Par la suite, nous montrerons (cf. § 1.31) qu'en dépit de ce que les pseudo-variables y_{ij} ne sont pas réparties normalement autour de la valeur (hypothétique) de la régression — ce qui est l'une des conditions d'applicabilité pour le test F — le rapport :

$$(120) \quad \frac{(N - p - 1) \sum_{i,j} (Y_{ij} - \bar{y})^2}{p \sum_{i,j} (Y_{ij} - y_{ij})^2} = \frac{N - p - 1}{p} \cdot \frac{r_p^2}{1 - r_p^2}$$

suit une répartition $F(p; N - p - 1)$. Il s'agira alors de préciser le sens d'un tel test, du moment que nous ne sommes pas en présence d'une régression au sens habituel.

TABLEAU 4

Variabilité	Degrés de liberté	Somme des carrés
régression	p	$\sum_{i,j} (Y_{ij} - \bar{y})^2 = \eta \sum_k b_k d_k = \eta r_p^2$
résiduelle	$N - p - 1$	$\sum_{i,j} (Y_{ij} - y_{ij})^2 = \eta (1 - \sum_k b_k d_k) = \eta (1 - r_p^2)$
totale	$N - 1$	$\sum_{i,j} (y_{ij} - \bar{y})^2 = \quad = \eta$

1.3 LA DISTANCE ENTRE LES ÉCHANTILLONS

1.31 Test de la distance entre deux échantillons.

En partant des relations:

$$(121) \quad \sum_k b_k S''_{kl} = \eta d_l (1 - \sum_k b_k d_k) \\ = \eta d_l (1 - r_p^2) \quad (l = 1, \dots, p) \quad (\text{cf. eq. 109 et 115}),$$

nous pouvons écrire:

$$(122) \quad n \sum_k b_k w_{kl} = \eta d_l (1 - r_p^2), \quad \text{où:}$$

$$(123) \quad w_{kl} = S''_{kl}/n, \\ n = N_1 + N_2 - 2.$$

Si $[w^{kl}]$ désigne la matrice inverse de $[w_{kl}]$, le coefficient b_k sera donné par:

$$(124) \quad nb_k = \eta (1 - r_p^2) \sum_l w^{kl} d_l. \quad \text{Posons maintenant:}$$

$$(125) \quad \xi_l = \sqrt{\eta} d_l = \frac{\bar{x}_{1l} - \bar{x}_{2l}}{\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \quad (l = 1, \dots, p). \quad \text{Donc:}$$

$$(126) \quad nb_k = \sqrt{\eta} (1 - r_p^2) \sum_l w^{kl} \xi_l \quad (k = 1, \dots, p).$$

En multipliant par $d_k = \xi_k/\sqrt{\eta}$ et en additionnant les p équations membre par membre, on obtient:

$$(127) \quad n \sum_k b_k d_k = nr_p^2 \\ = (1 - r_p^2) \sum_{k,l} w^{kl} \xi_k \xi_l.$$

Introduisons la variable T_p définie par la relation:

$$(128) \quad nT_p = \sum_{k,l} w^{kl} \xi_k \xi_l \quad (\text{cf. eq. 36}), \quad \text{de sorte que:}$$

$$(129) \quad nr_p^2 = n(1 - r_p^2) T_p, \quad \text{c'est-à-dire:}$$

$$(130) \quad r_p^2 = T_p/(1 + T_p), \quad \text{ou bien:}$$

$$(131) \quad T_p = r_p^2 / (1 - r_p^2) .$$

Dans les hypothèses du présent problème, on a alors :

$$(N - p - 1) T_p / p = F(p; N - p - 1) , \\ N = N_1 + N_2 . \quad (\text{Cf. eq. 38}).$$

Ce n'est autre chose que le résultat de HOTELLING, ou celui de FISHER, si l'on veut bien se souvenir de la relation entre la répartition de F et celle de z (cf. § 1.14) (HOTELLING, 1931, p. 377; FISHER, 1938, p. 378).

Considérons ensuite le rapport des deux carrés moyens (= sommes des carrés divisées par les degrés de liberté respectifs) supérieurs du tableau 4 (§ 1.23) :

$$(132) \quad \frac{\sum_{i,j} (Y_{ij} - \bar{y})^2 / p}{\sum_{i,j} (Y_{ij} - y_{ij})^2 / (N - p - 1)} = \frac{N - p - 1}{p} \cdot \frac{r_p^2}{1 - r_p^2} \\ = (N - p - 1) T_p / p .$$

On voit maintenant que ce rapport peut être testé par la répartition de F comme dans le cas d'une régression au sens ordinaire. L'hypothèse ici testée concerne les p écarts $d_k = \bar{x}_{1k} - \bar{x}_{2k}$; plus précisément, l'hypothèse-zéro est :

$$E(d_k) = 0 \quad \text{pour tout } k = 1, \dots, p .$$

On teste ainsi la qualité d'une discrimination grâce à une analyse de variance associée à la régression formelle d'une pseudo-variable y_{ij} en x_{ijk} .

En vertu de la relation $T_p = \nu D_p^2$ (eq. 45), on peut encore écrire :

$$(133) \quad (N - p - 1) T_p / p = (N - p - 1) \nu D_p^2 / p \\ = F(p; N - p - 1) , \quad \text{ou bien :}$$

$$(134) \quad \frac{(N - p - 1) N_1 N_2}{pnN} D_p^2 = F(p; N - p - 1) ,$$

$$N = N_1 + N_2 ; \quad n = N - 2 \quad (\text{RAO, 1952, p. 247}),$$

qui est le même test exprimé en termes de distance généralisée.

TABLEAU 5

Variabilité	Degrés de liberté	Somme des carrés	Carrés moyens	F
entre échantillons	p	S'_H	S'_H/p	$S'_H(N - p - 1)/pS''_H$
résiduelle	$N - p - 1$	S''_H	$S''_H/(N - p - 1)$	...
totale	$N - 1$	S_H	...	...

Remarquons enfin que ce test peut encore être conçu comme formellement déduit d'une analyse de variance entre échantillons / à l'intérieur des échantillons, établie à partir des valeurs individuelles H_{ij} du discriminant en question (tabl. 5). Car en partant de l'analyse de variance du tableau 2 (§ 1.12), on peut observer qu'en effet :

$$S'_H = \eta \left\{ \sum_k h_k d_k \right\}^2 \quad (\text{eq. 84}),$$

$$S_H'' = \sum_{k,l} h_k h_l S_{kl}'' \quad (\text{tabl. 2, § 1.12})$$

puisque $[S_{kl}''] [h_k] = [d_l]$ est le système d'équations déterminant les coefficients h_k . Or :

$$(135) \quad \sum_k h_k d_k = (\sum_k b_k d_k) / \eta (1 - r_p^2) \quad (\text{selon eq. 116})$$

$$= \frac{r_p^2}{1 - r_p^2} \eta (1 - r_p^2)$$

$$= D_p^2 / n \quad (\text{selon eq. 45 et 132}),$$

de sorte que finalement :

$$(136) \quad \frac{S_H' / p}{S_H'' / (N - p - 1)} = \frac{\eta D_p^4 / p n^2}{D_p^2 / n (N - p - 1)}$$

$$= (N - p - 1) \eta D_p^2 / p$$

$$= F(f_1; f_2)$$

comme précédemment (cf. FISHER, 1936, p. 179). Ceci montre, en particulier, que les degrés de liberté appropriés, dans le tableau 5, sont respectivement $f_1 = p$; $f_2 = N - p - 1$.

Il convient de citer, à cet endroit, les tables de BOSE pour le test d'hypothèse concernant la distance Δ_p^2 dans l'ensemble de base, en particulier $\Delta_p^2 = 0$, ainsi que ceux de CHOWDHURY indiquant les intervalles de confiance (*confidence intervals*) pour une valeur calculée de la distance D_p^2 (BOSE, 1946-48; p. 245; CHOWDHURY, 1954, p. 77).

1.32 Cas particulier de trois échantillons.

Le test précédent concerne la signification d'une discrimination sous l'hypothèse :

$$E(d_k) = 0 \quad \text{pour tout } k = 1, \dots, p,$$

$$d_k = \bar{x}_{1k} - \bar{x}_{2k}.$$

En d'autres termes, le test porte en fin de compte sur la distance D_p^2 entre deux échantillons. Lorsque ce test a mis en évidence une distance significative, on peut se demander quelle est la position, relativement à ces deux échantillons, d'un troisième échantillon resté inconsideré jusqu'ici. On cherchera à savoir, plus exactement, si ce troisième échantillon est plus proche du premier que du second, cette proximité étant mesurée par la distance généralisée.

Soit x_{ijk} ($i = 1, 2$ ou 3 ; $j = 1, \dots, N_i$; $k = 1, \dots, p$) l'ensemble des mesures de ces trois échantillons dans lesquels nous supposons l'égalité des variances et covariances. L'estimation de la matrice de dispersion $[w_{kl}]$ sera donnée par :

$$(137) \quad n w_{kl} = \sum_{i,j} (x_{ijk} - \bar{x}_{ik}) (x_{ijl} - \bar{x}_{il}),$$

$$n = N_1 + N_2 + N_3 - 3. \quad \text{Posons :}$$

$$(138) \quad \left. \begin{aligned} d'_k &= \bar{x}_{3k} - \bar{x}_{1k} \\ d''_k &= \bar{x}_{3k} - \bar{x}_{2k} \end{aligned} \right\} \quad (k = 1, \dots, p).$$

$$(139) \quad e_k = d'_k + d''_k$$

La matrice de dispersion (estimée) des e_k sera :

$$(140) \quad [w_{e_{kl}}] = (1/\psi) [w_{kl}], \quad \text{avec :}$$

$$(141) \quad 1/\psi = 1/N_1 + 1/N_2 + 4/N_3; \quad \text{degrés de liberté : } n.$$

Lorsqu'on pose :

$$(142) \quad E(d'_k) = E(-d''_k), \quad \text{ce qui revient à dire que :}$$

$$(143) \quad E(e_k) = 0, \quad \text{ou bien:}$$

$$(144) \quad 2E(x_{3jk}) - E(x_{1jk}) - E(x_{2jk}) = 0 \quad (k = 1, \dots, p), \text{ et en posant:}$$

$$(145) \quad nT_p = \sum_{k,l} w_e^{kl} e_k e_l, \quad \text{on sait que (cf. RAO, 1946, p. 413):}$$

$$(146) \quad (n - p + 1) T_p / p = F(p; n - p + 1), \\ n = N_1 + N_2 + N_3 - 3.$$

1.33 La classification d'une valeur individuelle.

Le test précédent peut être utilisé dans le cas où $N_3 = 1$. Il concerne alors le problème de situer une unique observation p -dimensionnelle par rapport à deux échantillons de N_1 et N_2 observations respectivement (cf. RAO, 1946, p. 413. Voir également WALD, 1944, p. 145). Soit donc $x_{31}, \dots, x_{3k}, \dots, x_{3p}$ cette observation; puisqu'elle est unique dans l'échantillon $i = 3$, on peut omettre l'indice j .

La supposition $N_3 = 1$ entraîne les simplifications suivantes:

$$(147) \quad nw_{kl} = \sum_{i=1}^2 \sum_{j=1}^{N_i} (x_{ijk} - \bar{x}_{ik})(x_{ijl} - \bar{x}_{il}), \\ n = N - 3 = N_1 + N_2 - 2. \quad \text{En effet, lorsque } i = 3:$$

$$(148) \quad \sum_{j=1}^{N_i} (x_{ijk} - \bar{x}_{ik})(x_{ijl} - \bar{x}_{il}) = 0$$

parce que $\bar{x}_{3k} = x_{3k}$ quel que soit $k = 1, \dots, p$. De cette façon:

$$(149) \quad \{ \psi(n - p + 1)/np \} \sum_{k,l} w^{kl} (2x_{3k} - \bar{x}_{1k} - \bar{x}_{2k})(2x_{3l} - \bar{x}_{1l} - \bar{x}_{2l}) = F(f_1; f_2),$$

ou bien finalement:

$$(150) \quad \{ \psi(n - p + 1)/np \} \sum_{k,l} w^{kl} e_k e_l = F(f_1; f_2), \quad \text{avec:}$$

$$(151) \quad e_k = 2x_{3k} - \bar{x}_{1k} - \bar{x}_{2k},$$

$$(152) \quad \psi = N_1 N_2 / (N_1 + N_2 + 4N_1 N_2),$$

$$f_1 = p$$

$$f_2 = N - p - 2 = N_1 + N_2 - p - 1.$$

$[w^{kl}]$ est l'inverse de la matrice $[w_{kl}]$ introduite ci-dessus (eq. 147); c'est la matrice de dispersion à l'intérieur des deux échantillons $i = 1$ et $i = 2$, et ayant $n = N_1 + N_2 - 2$ degrés de liberté.

Une valeur F significative indiquerait que l'observation x_{3k} peut être attribuée à l'un des deux échantillons, à cause de l'inégalité globalement significative des différences:

$$(153) \quad \left. \begin{aligned} d'_k &= x_{3k} - \bar{x}_{1k} \\ d''_k &= x_{3k} - \bar{x}_{2k} \end{aligned} \right\} \quad (k = 1, \dots, p).$$

Si, par contre, la valeur F est non-significative, la classification de l'observation x_{3k} reste indéterminée.

1.4 COMPARAISONS DE FONCTIONS DISCRIMINANTES

1.41 Précision d'une fonction discriminante.

Les coefficients h_k d'un discriminant $H = \sum_k h_k x_k$ ne sont déterminés qu'à une constante de proportionnalité λ ($\neq 0$) près, c'est-à-dire que seul le rapport h_k/h_l de deux coefficients ($k \neq l$) est univoquement déterminable (cf. § 1.21). Un coefficient d'un discri-

minant ne se comporte donc pas comme l'estimateur d'un certain paramètre défini dans l'ensemble de base. En particulier, aucune mesure de dispersion simple ne peut être définie pour un coefficient déterminé, et dès lors il n'est pas très évident comment on pourrait construire un test de signification pour chacun des coefficients séparément tel qu'il existe pour les coefficients d'une régression (linéaire) multiple par exemple.

Ce que l'on peut faire, par contre, c'est de définir empiriquement une fonction discriminante:

$$(154) \quad G = \sum_k g_k x_k, \quad \text{c'est-à-dire:}$$

$$G_{ij} = \sum_k g_k x_{ijk},$$

en choisissant *a priori*, disons, la valeur des coefficients g_k . On comparerait le comportement d'une telle fonction empirique à celui du (meilleur) discriminant calculé d'après les données observées, à savoir:

$$(155) \quad H = \sum_k h_k x_k, \quad \text{c'est-à-dire:}$$

$$H_{ij} = \sum_k h_k x_{ijk}.$$

Le choix des coefficients g_k serait fait, par exemple, à la lumière de certaines considérations relevant du domaine particulier d'investigation auquel on désirerait appliquer l'analyse discriminante.

Une telle comparaison peut être basée, selon la suggestion de FISHER, sur la corrélation entre les valeurs G_{ij} et H_{ij} , corrélation calculée à l'intérieur des échantillons. Si cette corrélation diminue lorsqu'on fait varier, dans un sens déterminé, les coefficients g_k , on considérera le discriminant G comme étant de plus en plus en contradiction avec les données observées, en particulier avec le meilleur discriminant H déterminé par elles. Indiquer la précision du meilleur discriminant H revient ainsi à assigner une limite de signification à cette corrélation (FISHER, 1940, p. 423).

Soit donc la régression multiple:

$$(156) \quad Y_{ij} = \sum_k b_k x_{ijk}$$

des pseudo-variables y_{ij} en les observations x_{ijk} ($i = 1$ ou 2 ; $j = 1, \dots, N_i$; $k = 1, \dots, p$), avec:

$$(157) \quad y_{ij} = N_2/N \quad \text{pour } i = 1$$

$$y_{ij} = -N_1/N \quad \text{pour } i = 2; N = N_1 + N_2.$$

Les valeurs x_{ijk} sont maintenant supposées mesurées à partir de leur moyenne générale; c'est-à-dire que $\bar{x}_k = 0$ quel que soit k . De même $\bar{y} = 0$ à cause des définitions mêmes. Les coefficients b_k sont alors donnés par les équations:

$$(158) \quad \sum_k b_k S_{kl} = S_{ly} \quad (l = 1, \dots, p), \text{ avec par définition:}$$

$$(159) \quad S_{kl} = \sum_{i,j} x_{ijk} x_{ijl},$$

$$(160) \quad S_{ly} = \sum_{i,j} x_{ijl} y_{ij}$$

= ηd_l en tenant compte des conventions:

$$d_l = \bar{x}_{1l} - \bar{x}_{2l} \quad (l = 1, \dots, p),$$

$$\eta = N_1 N_2 / (N_1 + N_2).$$

Comme précédemment, on peut établir une analyse de variance (tabl. 6). On a de nouveau :

$$\begin{aligned}
 (161) \quad r_p^2 &= \sum_k b_k S_{ky} / S_{yy} \\
 &= \eta \sum_k b_k d_k / \eta \\
 &= \sum_k b_k d_k \quad (\text{cf. cq. 119}).
 \end{aligned}$$

TABLEAU 6

Variabilité	Degrés de liberté	Somme des carrés
régression	p	$\sum_{i,j} Y_{ij}^2 = r_p^2 \eta$
résiduelle	$N - p - 1$	$\sum_{i,j} (Y_{ij} - y_{ij})^2 = (1 - r_p^2) \eta$
totale	$N - 1$	$\sum_{i,j} y_{ij}^2 = \eta$

Mais les coefficients b_k , à une constante multiplicative près, peuvent être pris pour coefficients d'un discriminant (cf. § 1.22); la régression $Y = \sum_k b_k x_k$ peut donc être considérée comme une fonction discriminante, et on peut établir l'analyse de variance entre les échantillons/ à l'intérieur des échantillons, pour la variable Y_{ij} (tabl. 7, dernière colonne), en tenant compte pour cela des relations qui vont suivre :

$$(162) \quad \bar{Y} = 0, \quad \text{car} \quad \bar{Y} = \sum_k b_k \bar{x}_k, \quad \text{et} \quad \bar{x}_k = 0;$$

$$(163) \quad \sum_{i,j} (Y_{ij} - \bar{Y})^2 = \sum_{i,j} Y_{ij}^2 = \eta r_p^2 \quad (\text{cf. § 1.23});$$

$$\begin{aligned}
 (164) \quad \sum_{i,j} (\bar{Y}_i - \bar{Y})^2 &= \sum_{i,j} \bar{Y}_i^2 \\
 &= \eta (\bar{Y}_1 - \bar{Y}_2)^2 \\
 &= \eta r_p^4 \quad \text{parce que d'autre part:}
 \end{aligned}$$

$$(165) \quad \bar{Y}_1 - \bar{Y}_2 = \sum_k b_k d_k = r_p^2. \quad \text{Par conséquent:}$$

$$(166) \quad \sum_{i,j} (\bar{Y}_i - Y_{ij})^2 = \eta r_p^2 - \eta r_p^4 = \eta r_p^2 (1 - r_p^2).$$

Soit maintenant $G_{ij} = \sum_k g_k x_{ijk}$ la fonction empirique; supposons, pour fixer les idées, que ses coefficients aient été donnés par avance. On a les relations :

$$(167) \quad \bar{G} = \sum_k g_k \bar{x}_k = 0 \quad \text{puisque} \quad \bar{x}_k = 0;$$

$$(168) \quad c^2 = \sum_{i,j} G_{ij}^2 \quad (\text{convention d'écriture}).$$

Soit r la corrélation (simple) totale entre Y_{ij} et G_{ij} , donc par définition (cf. § 1.12):

$$(169) \quad r = \text{Cov} (G; Y) \{ \text{Var} (G) \text{Var} (Y) \}^{-\frac{1}{2}},$$

la variance et la covariance étant entendues prises sur la réunion des deux échantillons. On a alors :

$$(170) \quad \sum_{i,j} G_{ij} Y_{ij} = r \{ \text{Var} (G) \text{Var} (Y) \}^{\frac{1}{2}} = r c r_p \sqrt{\eta}.$$

On a d'autre part :

$$\begin{aligned}
 (171) \quad \sum_{i,j} Y_{ij} G_{ij} &= \sum_{i,j} \{ (\sum_k b_k x_{ijk}) (\sum_l g_l x_{ijl}) \} \\
 &= \sum_l g_l \sum_k b_k s_{kl} \\
 &= \eta \sum_k g_k d_k, \quad \text{et aussi:}
 \end{aligned}$$

$$\begin{aligned}
 (172) \quad \sum_{i,j} y_{ij} G_{ij} &= (N_2/N) \sum_{j=1}^{N_1} \sum_l g_l x_{1jl} + (-N_1/N) \sum_{j=1}^{N_2} \sum_l g_l x_{2jl} \\
 &= \eta \sum_l g_l (\bar{x}_{1l} - \bar{x}_{2l}) \\
 &= \eta \sum_k g_k d_k, \quad \text{d'où:}
 \end{aligned}$$

$$\begin{aligned}
 (173) \quad \sum_{i,j} y_{ij} G_{ij} &= \sum_{i,j} Y_{ij} G_{ij} \\
 &= \sqrt{\eta} r r_p c, \quad \text{c'est-à-dire:}
 \end{aligned}$$

$$(174) \quad \sum_{i,j} (y_{ij} - Y_{ij}) G_{ij} = 0. \quad \text{Enfin:}$$

$$\begin{aligned}
 (175) \quad \eta (\bar{G}_1 - \bar{G}_2) &= \eta \left\{ \sum_{j=1}^{N_1} G_{1j}/N_1 - \sum_{j=1}^{N_2} G_{2j}/N_2 \right\} \\
 &= (N_2/N) \sum_{j=1}^{N_1} G_{1j} + (-N_1/N) \sum_{j=1}^{N_2} G_{2j} \\
 &= \sum_{j=1}^{N_1} y_{1j} G_{1j} + \sum_{j=1}^{N_2} y_{2j} G_{2j} \\
 &= \sum_{i,j} y_{ij} G_{ij} \\
 &= \sqrt{\eta} r r_p c; \quad \text{et également:}
 \end{aligned}$$

$$\begin{aligned}
 (176) \quad \sum_{i,j} (\bar{G}_i - \bar{G})^2 &= \eta (\bar{G}_1 - \bar{G}_2)^2 \\
 &= r^2 r_p^2 c^2. \quad \text{Finalement:}
 \end{aligned}$$

$$\begin{aligned}
 (177) \quad \sum_{i,j} \bar{G}_i \bar{Y}_i &= \sum_{i,j} (\bar{G}_i - \bar{G}) (\bar{Y}_i - \bar{Y}) \\
 &= \eta (\bar{G}_1 - \bar{G}_2) (\bar{Y}_1 - \bar{Y}_2) \\
 &= \sqrt{\eta} r r_p c \cdot r_p^2 = \sqrt{\eta} r r_p^3 c,
 \end{aligned}$$

en utilisant la relation, analogue aux équations 9 ou 12, entre la moyenne générale et la moyenne dans l'échantillon. Toutes ces relations sont résumées dans l'analyse de covariance du tableau 7.

TABLEAU 7

Variabilité	Somme des carrés ou des produits		
	(GG)	(GY)	(YY)
entre échantillons	$r^2 r_p^2 c^2$	$\sqrt{\eta} r r_p^3 c$	ηr_p^4
à l'intérieur	$(1 - r^2 r_p^2) c^2$	$\sqrt{\eta} r r_p (1 - r_p^2) c$	$\eta r_p^2 (1 - r_p^2)$
totale	c^2	$\sqrt{\eta} r r_p c$	ηr_p^2

Introduisons maintenant la corrélation (simple) r'' à l'intérieur des échantillons obtenue en prenant dans l'expression définissant r (eq. 169) les variances et la covariance à l'intérieur des échantillons, et non plus dans le total:

$$(178) \quad r''^2 = \{ \sqrt{\eta} r r_p (1 - r_p^2) c \}^2 / \{ (1 - r^2 r_p^2) c^2 \cdot \eta r_p^2 (1 - r_p^2) \} \\ = r^2 (1 - r_p^2) / (1 - r^2 r_p^2), \quad \text{ou bien:}$$

$$(179) \quad r^2 = r''^2 / \{ 1 - r_p^2 (1 - r''^2) \} \quad \text{en résolvant par rapport à } r.$$

Considérons ensuite la régression linéaire simple de y_{ij} en G_{ij} :

$$(180) \quad \gamma_{ij} = b_{y.G} G_{ij}, \quad \text{avec:}$$

$\bar{G} = 0$ (eq. 167), et $\bar{y} = 0$ (eq. 157; définitions). On aura:

$$(181) \quad b_{y.G} = \frac{\sum_{i,j} y_{ij} G_{ij} / \sum_{i,j} G_{ij}^2}{\sqrt{\eta} r r_p c / c^2} \\ = \sqrt{\eta} r r_p / c.$$

La variabilité totale $\sum_{i,j} y_{ij}^2$ peut être subdivisée selon l'analyse de la variance du tableau 8.

TABLEAU 8

Variabilité	DL	Somme des carrés
sur la régression	1	$\sum_{i,j} \gamma_{ij}^2 = b_{y.G} \sum_{i,j} y_{ij} G_{ij} = \eta r^2 r_p^2$ $= \eta r''^2 r_p^2 / \{ 1 - r_p^2 (1 - r''^2) \}$
autour de la régression	$N - 2$	$\sum_{i,j} (y_{ij} - \gamma_{ij})^2 = \eta (1 - r^2 r_p^2)$ $= \eta (1 - r_p^2) / \{ 1 - r_p^2 (1 - r''^2) \}$
totale	$N - 1$	$\sum_{i,j} y_{ij}^2 = \eta$

Remarquons que la somme des carrés résiduelle $\sum_{i,j} (y_{ij} - \gamma_{ij})^2$ peut, à son tour, être partagée en deux parties:

$$(182) \quad \sum_{i,j} (y_{ij} - \gamma_{ij})^2 = \sum_{i,j} (Y_{ij} - \gamma_{ij} + y_{ij} - Y_{ij})^2 \\ = \sum_{i,j} (Y_{ij} - \gamma_{ij})^2 + \sum_{i,j} (y_{ij} - Y_{ij})^2 \quad \text{parce que:}$$

$$(183) \quad \sum_{i,j} (Y_{ij} - \gamma_{ij}) (y_{ij} - Y_{ij}) = 0;$$

cette dernière relation provient de ce que:

$$(184) \quad \sum_{i,j} Y_{ij} (y_{ij} - Y_{ij}) = \sum_{i,j} y_{ij} Y_{ij} - \sum_{i,j} Y_{ij}^2 \\ = \sum_{i,j} y_{ij} \sum_k b_k x_{ijk} - \eta r_p^2 \\ = \eta \{ \sum_k b_k d_k - r_p^2 \} = 0, \quad \text{et de ce que:}$$

$$(185) \quad \sum_{i,j} \gamma_{ij} (y_{ij} - Y_{ij}) = b_{y.G} \sum_{i,j} G_{ij} (y_{ij} - Y_{ij}) = 0 \quad (\text{cf. eq. 174}).$$

C'est ainsi qu'on aboutit à l'analyse de variance du tableau 9.

TABLEAU 9

Variabilité	Degrés de liberté	Somme des carrés
entre Y et γ	$p - 1$	$\sum_{i,j} (Y_{ij} - \gamma_{ij})^2 = \eta r_p^2 (1 - r_p^2) / \{ 1 - r_p^2 (1 - r'^2) \}$
résiduelle	$N - p - 1$	$\sum_{i,j} (y_{ij} - Y_{ij})^2 = \eta (1 - r_p^2)$
autour de la régression	$N - 2$	$\sum_{i,j} (y_{ij} - \gamma_{ij})^2 = \eta (1 - r_p^2) / \{ 1 - r_p^2 (1 - r'^2) \}$

On peut donner à la somme des carrés $\sum_{i,j} (Y_{ij} - \gamma_{ij})^2$ l'interprétation suivante. Soit:

$$(186) \quad \Gamma_{ij} = b_{Y.G} G_{ij}$$

la régression linéaire simple de Y_{ij} en G_{ij} , avec les propriétés:

$$\bar{Y} = 0 \quad (\text{eq. 162}) \quad \text{et} \quad \bar{G} = 0 \quad (\text{eq. 167}), \quad \text{et:}$$

$$(187) \quad b_{Y.G} = \frac{\sum_{i,j} G_{ij} Y_{ij}}{\sum_{i,j} G_{ij}^2} \\ = \sqrt{\eta} r r_p / c.$$

Ce coefficient $b_{Y.G}$ est du reste identique au coefficient $b_{y.G}$ (eq. 181). On peut associer à cette régression Γ_{ij} l'analyse de variance reproduite au tableau 10, et qui concerne les valeurs de la variable « dépendante » Y_{ij} . Cette analyse de variance confirme, en particulier, que la somme des carrés $\sum_{i,j} (Y_{ij} - \Gamma_{ij})^2$ a $p - 1$ degrés de liberté. Elle montre par ailleurs que:

$$(188) \quad \sum_{i,j} (Y_{ij} - \Gamma_{ij})^2 = \sum_{i,j} (Y_{ij} - \gamma_{ij})^2.$$

TABLEAU 10

Variabilité	DL	Somme des carrés
sur la régression	1	$\sum_{i,j} \Gamma_{ij}^2 = b_{Y.G}^2 \sum_{i,j} G_{ij}^2 = r^2 r_p^2 \eta = \eta r_p^2 r'^2 / \{ 1 - r_p^2 (1 - r'^2) \}$
autour de la régression	$p - 1$	$\sum_{i,j} (Y_{ij} - \Gamma_{ij})^2 = \frac{\eta r_p^2 (1 - r^2)}{\eta r_p^2 (1 - r_p^2) (1 - r'^2) / \{ 1 - r_p^2 (1 - r'^2) \}}$
totale	p	$\sum_{i,j} Y_{ij}^2 = \eta r_p^2$

Considérons maintenant le rapport suivant:

$$(189) \quad r_d^2 = \frac{\sum_{i,j} (Y_{ij} - \gamma_{ij})^2}{\sum_{i,j} (y_{ij} - \gamma_{ij})^2} \quad (\text{voir tabl. 9}),$$

quotient de deux sommes de carrés ayant la signification suivante:

$$\sum_{i,j} (Y_{ij} - \gamma_{ij})^2 \quad \text{est la somme des carrés sur la régression } Y \text{ lorsqu'on a éliminé la portion attribuable aux variations de } G; \text{ elle a } p - 1 \text{ degrés de liberté;}$$

$\sum_{i,j} (y_{ij} - \gamma_{ij})^2$ est la somme des carrés de la variable y lorsqu'on a tenu compte de l'influence de la corrélation entre y et G ; elle a $N - 2$ degrés de liberté.

Le quotient r_d^2 peut donc être considéré comme coefficient de corrélation débarrassé de l'influence de la variable G_{ij} ; il peut être testé comme précédemment :

$$(190) \quad \frac{r_d^2}{1 - r_d^2} \cdot \frac{N - p - 1}{p - 1} = F(f_1; f_2),$$

$$f_1 = p - 1; f_2 = N - 2 - (p - 1) = N - p - 1.$$

En substituant, on trouve finalement pour test :

$$(191) \quad \frac{r_p^2 (1 - r''^2)}{1 - r_p^2 (1 - r''^2)} \cdot \frac{N - p - 1}{p - 1} = F(p - 1; N - p - 1);$$

c'est le rapport des deux carrés moyens supérieurs dans l'analyse de variance du tableau 9. On observe que c'est — formellement — le test de r_p^2 du paragraphe 1.31 (eq. 132) où l'on aurait posé :

$$(192) \quad r_d^2 = r_p^2 (1 - r''^2),$$

et réduit les p degrés de liberté d'une unité. On conviendra de considérer une fonction G_{ij} empirique (désignée *a priori*) comme non contradictoire avec les données observées chaque fois que le coefficient r_d^2 ainsi testé n'est pas significatif. En d'autres termes, on rejettera toute fonction G proposée par avance, si sa corrélation r'' (à l'intérieur des échantillons) avec le meilleur discriminant H déterminé par les données observées, est suffisamment petite pour que le coefficient r_d^2 soit significatif (cf. FISHER, 1958, p. 287).

A toutes fins utiles, mentionnons encore les travaux de BARTLETT sur la précision des coefficients d'une fonction discriminante (BARTLETT, 1939a, p. 169), et plus généralement sur les tests de signification concernant des fonctions discriminantes (BARTLETT, 1938, p. 33; BARTLETT, 1939b, p. 180).

1.42 Comparaison avec une fonction empirique.

On peut donner du test précédent (eq. 191) une interprétation plus directe en utilisant le concept de distance généralisée associée à une fonction empirique, c'est-à-dire désignée *a priori*, par exemple. Reprenons la relation 191 :

$$(193) \quad \frac{N - p - 1}{p - 1} \cdot \frac{r_p^2 (1 - r''^2)}{1 - r_p^2 (1 - r''^2)} = F(p - 1; N - p - 1),$$

où r'' est la corrélation simple, à l'intérieur des deux échantillons, entre le meilleur discriminant H déterminé par les données observées, et la fonction empirique G en question :

$$(194) \quad r''^2 = \text{cov}^2 (G; H) / \{ \text{var} (G) \text{var} (H) \},$$

où $\text{cov} (...)$ et $\text{var} (...)$ sont entendues à l'intérieur des échantillons par définition. On a aussi :

$$(195) \quad n \text{ cov} (G; H) = \sum_{i,j} (G_{ij} - \bar{G}_i) (H_{ij} - \bar{H}_i)$$

$$= \sum_{i,j} (G_{ij} - \bar{G}) (H_{ij} - \bar{H}) - \sum_{i,j} (\bar{G}_i - \bar{G}) (\bar{H}_i - \bar{H}),$$

en vertu d'une analyse de covariance entre / à l'intérieur des deux échantillons, ayant la structure du tableau 2 (§ 1.12). En fixant la fonction H par les coefficients b_k déterminés par la pseudo-régression multiple Y_{ij} (eq. 158), on obtient :

$$\begin{aligned}
 (196) \quad \sum_{i,j} (G_{ij} - \bar{G}) (H_{ij} - \bar{H}) &= \sum_{i,j} \sum_k b_k \sum_l g_l (x_{ijk} - \bar{x}_k) (x_{ijl} - \bar{x}_l) \\
 &= \sum_l g_l \sum_k b_k S_{kl} \\
 &= \sum_l g_l \eta d_l \\
 &= \eta (\bar{G}_1 - \bar{G}_2) ;
 \end{aligned}$$

$$\begin{aligned}
 (197) \quad \sum_{i,j} (\bar{G}_i - \bar{G}) (\bar{H}_i - \bar{H}) &= \eta (\bar{G}_1 - \bar{G}_2) \sum_k b_k d_k \\
 &= \eta (\bar{G}_1 - \bar{G}_2) (\bar{H}_1 - \bar{H}_2) \\
 &= \eta r_p^2 (\bar{G}_1 - \bar{G}_2) ,
 \end{aligned}$$

où $r_p^2 = \sum_k b_k d_k = \bar{H}_1 - \bar{H}_2$ (eq. 119); c'est ainsi que:

$$(198) \quad n \operatorname{cov} (G; H) = \eta (1 - r_p^2) (\bar{G}_1 - \bar{G}_2) . \quad \text{D'autre part:}$$

$$(199) \quad n \operatorname{var} (H) = \sum_{i,j} (H_{ij} - \bar{H}_i)^2 = \eta r_p^2 (1 - r_p^2) \quad (\text{tabl. 7, § 1.41}), \text{ d'où:}$$

$$\begin{aligned}
 (200) \quad r'^2 &= \{ \eta (1 - r_p^2) (\bar{G}_1 - \bar{G}_2) \}^2 / \{ \eta r_p^2 (1 - r_p^2) n \operatorname{var} (G) \} \\
 &= \frac{\eta (\bar{G}_1 - \bar{G}_2)^2}{n \operatorname{var} (G)} \cdot \frac{1 - r_p^2}{r_p^2} .
 \end{aligned}$$

Les coefficients h_k du meilleur discriminant H sont donnés, dans le cas de la détermination par la méthode de l'extremum, par les équations:

$$(201) \quad \sum_k h_k w_{kl} = d_l \quad (l = 1, \dots, p), \quad \text{avec:}$$

$$(202) \quad n w_{kl} = \sum_{i,j} (x_{ijk} - x_{ik}) (x_{ijl} - x_{il}) ; \quad \text{done:}$$

$$(203) \quad h_l = \sum_k w^{kl} d_k \quad (l = 1, \dots, p), \text{ ainsi que:}$$

$$\begin{aligned}
 (204) \quad \sum_l h_l \sum_k h_k w_{kl} &= \sum_l d_l h_l \\
 &= \sum_{k,l} w^{kl} d_k d_l \quad (\text{cf. § 1.21}).
 \end{aligned}$$

Avec ces coefficients h_k , et compte tenu de la définition de la distance D_p^2 , on aura:

$$\begin{aligned}
 (205) \quad D_p^2 &= \sum_{k,l} w^{kl} d_k d_l \quad (\text{cf. eq. 32}) \\
 &= \{ \sum_l h_l d_l \}^2 / \sum_{k,l} h_k h_l w_{kl} \\
 &= (\bar{H}_1 - \bar{H}_2)^2 / \operatorname{var} (H) .
 \end{aligned}$$

Définissons par analogie, pour une fonction G :

$$(206) \quad D_G^2 = (\bar{G}_1 - \bar{G}_2)^2 / \operatorname{var} (G) , \quad \text{avec:}$$

$$(207) \quad n \operatorname{var} (G) = n \sum_{k,l} g_k g_l w_{kl} ;$$

en substituant dans la relation 200 ci-dessus, on obtiendra:

$$(208) \quad r'^2 = \nu D_G^2 (1 - r_p^2) / r_p^2 , \quad \text{avec:}$$

$$(209) \quad \nu = \eta / n .$$

D'autre part, $n T_p = n r_p^2 / (1 - r_p^2) = \eta D_p^2$ (eq. 131), ce qui donne:

$$(210) \quad r_p^2 = \nu D_p^2 / (1 + \nu D_p^2) , \quad \text{et aussi:}$$

$$(211) \quad 1 - r_p^2 = 1 / (1 + \nu D_p^2) . \quad \text{D'où:}$$

$$(212) \quad r''^2 = \nu D_G^2 / \nu D_p^2 \\ = D_G^2 / D_p^2 \quad \text{simplement. Posons maintenant:}$$

$$(213) \quad U = \frac{r_p^2 (1 - r''^2)}{1 - r_p^2 (1 - r''^2)} \\ = \frac{\nu D_p^2 (1 - D_G^2 / D_p^2)}{(1 + \nu D_p^2) - \nu D_p^2 (1 - D_G^2 / D_p^2)} \\ = \frac{\nu D_p^2 - \nu D_G^2}{1 + \nu D_G^2} + 1 - 1 \\ = \frac{1 + \nu D_p^2}{1 + \nu D_G^2} - 1, \quad \text{et finalement:}$$

$$(214) \quad U(N - p - 1) / (p - 1) = F(p - 1; N - p - 1).$$

On a ainsi le test de comparaison, basé sur les distances associées respectivement au meilleur discriminant H déterminé par les données observées, et une fonction G désignée par avance (empiriquement par exemple):

$$(215) \quad \frac{N - p - 1}{p - 1} \cdot \left\{ \frac{1 + \nu D_p^2}{1 + \nu D_G^2} - 1 \right\} = F(p - 1; N - p - 1), \quad \text{avec:} \\ N = N_1 + N_2; \quad \nu = N_1 N_2 / (N - 2) N \quad (\text{RAO, 1952, p. 249}).$$

Citons WILLIAMS indiquant une analyse de variance qui subdivise la variabilité totale en trois parties: une somme de carrés, entre échantillons, des valeurs prises par un discriminant hypothétique à coefficients désignés d'avance; une somme de carrés caractérisant les écarts entre ces coefficients désignés d'avance et ceux du discriminant estimé d'après les données; enfin une somme de carrés résiduelle (WILLIAMS, 1959, p. 179).

1.43 Comparaison de deux discriminations calculées.

Nous supposons avoir quatre échantillons d'observations $x_{ijk}: i = 1, \dots, 4; j = 1, \dots, N_i; k = 1, \dots, p; N = \sum_i N_i; n = N - 4$. Les variances et covariances seront supposées égales par hypothèse, de sorte qu'on a pour estimation $[w_{kl}]$ de la matrice de dispersion:

$$(216) \quad nw_{kl} = \sum_{i,j} (x_{ijk} - \bar{x}_{ik})(x_{ijl} - \bar{x}_{il}).$$

Posons maintenant:

$$(217) \quad \left. \begin{aligned} d'_k &= \bar{x}_{1k} - \bar{x}_{2k}, \\ d''_k &= \bar{x}_{3k} - \bar{x}_{4k}, \end{aligned} \right\} (k = 1, \dots, p)$$

$$(217a) \quad e_k = d'_k - d''_k.$$

La matrice de dispersion (estimée) des e_k sera donc:

$$(218) \quad [w_{ekl}] = (1/\psi) [w_{kl}] \quad \text{où:}$$

$$(219) \quad 1/\psi = \sum_i (1/N_i) \quad (\text{quadruple moyenne harmonique}). \quad \text{Calculons:}$$

$$(220) \quad nT_p = \sum_{k,l} w_e^{kl} d_k d_l \quad \text{où:}$$

$$(221) \quad [w_e^{kl}] = \psi [w^{kl}].$$

Lorsqu'on fait l'hypothèse suivante:

$$(222) \quad E(e_k) = 0 \quad (k = 1, \dots, p), \quad \text{c'est-à-dire:}$$

$$(223) \quad E(d'_k) = E(d''_k) \quad (k = 1, \dots, p),$$

on sait que $(n - p + 1) T_p/p$ est réparti comme $F(p; n - p + 1)$ (cf. § 1.13). Donc:

$$(224) \quad \left\{ \psi(n - p + 1)/np \right\} \sum_{k,l} w^{kl} (d'_k - d''_k) (d'_l - d''_l) = F(p; n - p + 1),$$

$$n = \sum_i (N_i - 1); \quad \psi = 1/\left\{ \sum_i (1/N_i) \right\}. \quad (\text{Cf. RAO, 1952, p. 251.})$$

Remarquons que la numérotation des échantillons n'a pas de signification propre: au lieu de comparer les écarts $\bar{x}_{1k} - \bar{x}_{2k}$ aux écarts $\bar{x}_{3k} - \bar{x}_{4k}$, on pourrait aussi les comparer aux écarts $\bar{x}_{4k} - \bar{x}_{3k} = -(\bar{x}_{3k} - \bar{x}_{4k})$, c'est-à-dire faire les hypothèses:

$$(225) \quad E(e'_k) = 0 \quad (k = 1, \dots, p), \text{ où:}$$

$$(226) \quad e'_k = d'_k + d''_k; \quad \text{autrement dit, on ferait l'hypothèse:}$$

$$(227) \quad E(d'_k) = E(-d''_k) \quad (k = 1, \dots, p).$$

On effectuerait alors le test avec la statistique:

$$(228) \quad \left\{ \psi(n - p + 1)/np \right\} \sum_{k,l} w^{kl} (d'_k + d''_k) (d'_l + d''_l),$$

répartie comme $F(p; n - p + 1)$, avec $n = \sum_i (N_i - 1)$, et $\psi = 1/\left\{ \sum_i (1/N_i) \right\}$.

Tester l'hypothèse $E(e_k) = 0$ sous l'une ou l'autre forme, revient à tester l'égalité de la discrimination entre le premier et le second, et celle entre le troisième et le quatrième échantillon, autrement dit, à tester l'égalité de deux distances généralisées calculées l'une entre le premier et le deuxième échantillon, l'autre entre le troisième et le quatrième échantillon.

1.5 SIMPLIFICATIONS DE FONCTIONS DISCRIMINANTES

1.51 L'inclusion de nouvelles variables.

Rappelons que si, dans la définition de la distance généralisée:

$$(229) \quad D_p^2 = \sum_{k,l} s^{kl} (\bar{x}_{1k} - \bar{x}_{2k}) (\bar{x}_{1l} - \bar{x}_{2l})$$

entre deux échantillons d'observations x_{ijk} ($i = 1$ ou 2 ; $j = 1, \dots, N_i$; $k = 1, \dots, p$) à dispersion estimée $[s_{kl}]$ avec $n = N_1 + N_2 - 2$ degrés de liberté, à savoir:

$$(230) \quad ns_{kl} = \sum_{i,j} (x_{ijk} - \bar{x}_{ik}) (x_{ijl} - \bar{x}_{il}),$$

l'on pose:

$$(231) \quad d_k = \bar{x}_{1k} - \bar{x}_{2k},$$

$$(232) \quad [w_{kl}] = (1/N_1 + 1/N_2) [s_{kl}],$$

on retombe sur la définition du rapport généralisé T_p , c'est-à-dire:

$$(233) \quad nT_p = \sum_{k,l} w^{kl} d_k d_l.$$

On a alors la relation:

$$(234) \quad T_p = \nu D_p^2,$$

$$\nu = N_1 N_2 / n (N_1 + N_2) \quad (\text{eq. 45}).$$

Adjoignons maintenant aux p variables considérées au départ, q nouvelles variables indépendantes d'elles. Soit T_{p+q} le rapport généralisé qui résulte de l'observation de l'en-

semble de ces $p + q$ variables, par opposition à T_p ne dépendant que des p anciennes variables. Sans égard aux hypothèses concernant les p premières variables, l'hypothèse:

$$(235) \quad E(d_{p+1}) = 0, E(d_{p+2}) = 0, \dots, E(d_{p+q}) = 0$$

implique que:

$$(236) \quad \frac{N_1 + N_2 - p - q - 1}{q} \cdot \left\{ \frac{1 + T_{p+q}}{1 + T_p} - 1 \right\} = F(q; N_1 + N_2 - p - q - 1) \quad (\text{cf. eq. 40}).$$

En utilisant la relation ci-dessus (eq. 234) entre les grandeurs D_p^2 et T_p , on aboutit à la formule:

$$(237) \quad \frac{n - p - q + 1}{q} \cdot \left\{ \frac{1 + \nu D_{p+q}^2}{1 + \nu D_p^2} - 1 \right\} = F(q; n - p - q + 1),$$

$$n = N_1 + N_2 - 2; \quad \nu = N_1 N_2 / n (N_1 + N_2), \quad (\text{RAO, 1952, p. 253}).$$

D_{p+q}^2 est la distance obtenue par la considération de l'ensemble des $p + q$ variables. On peut donc ainsi tester si l'adjonction, aux p variables utilisées antérieurement, de q variables supplémentaires entraîne une augmentation significative de la distance généralisée.

Lorsqu'on pose $p = 0$ et $D_0^2 = 0$ par convention naturelle, le test précédent prend la forme suivante:

$$(238) \quad \frac{n - q + 1}{q} \nu D_q^2 = F(q; n - q + 1)$$

qui n'est autre que le test d'une distance D_p^2 basée sur q variables (cf. eq. 134).

1.52 Importance, pour la discrimination, d'une variable déterminée.

La difficulté que l'on rencontre lorsqu'on examine l'importance de l'une déterminée des variables dans une discrimination provient, rappelons-le, de ce que les coefficients h_k de la fonction discriminante ne sont déterminables qu'à une constante de proportionnalité près (cf. §§ 1.21 et 1.41). Cela rend ainsi illusoire tout test, concernant l'un des coefficients h_k déterminé, qui utiliserait ce dernier autrement qu'à travers le rapport h_k/h_l ($k \neq l$) à l'un quelconque des autres coefficients.

L'une des méthodes (celle proposée et réalisée par FISHER; cf. §§ 1.41 et 1.42) de contourner cette difficulté consiste à faire intervenir une fonction discriminante désignée par avance, soit G , et d'étudier la corrélation entre G et le meilleur discriminant H déterminé par les données observées. A moins qu'on fixe tous les coefficients g_k du discriminant empirique G sauf un, en fonction des variations duquel on étudierait alors cette corrélation, ce test ne fournit cependant pas d'information sur l'influence directe d'une seule variable sur la discrimination.

Une autre méthode résulte immédiatement, par contre, de l'application du test du paragraphe précédent (§ 1.51). Si l'on a p variables et qu'on désire examiner l'effet, sur la discrimination entre deux échantillons, de l'inclusion d'une $(p + 1)$ -ième variable, on peut calculer les deux distances D_p^2 et D_{p+1}^2 , avec p et $p + 1$ variables respectivement. Puis on testerait selon la formule 237 dans laquelle on aurait posé, au préalable, $q = 1$:

$$(239) \quad (n - p) \left\{ \frac{1 + \nu D_{p+1}^2}{1 + \nu D_p^2} - 1 \right\} = F(1; n - p), \quad \text{avec:}$$

$$n = N_1 + N_2 - 2; \quad \nu = N_1 N_2 / n (N_1 + N_2).$$

Très exactement, ce test concerne l'augmentation de la distance généralisée due à l'adjonction de la variable x_{p+1} aux p autres variables (cf. RAO, 1952, p. 251).

On pourrait évidemment raisonner en sens inverse et comparer la distance D_p^2 à une distance D_{p-1}^2 qui ne tient plus compte que de $p - 1$ variables; nous sommes donc dans le cas où nous excluons une variable, et l'effet sera alors testé par la quantité:

$$(240) \quad (n - p + 1) \left\{ \frac{1 + \nu D_p^2}{1 + \nu D_{p-1}^2} - 1 \right\} = F(1; n - p + 1), \quad \text{avec:}$$

$$n = N_1 + N_2 - 2; \quad \nu = N_1 N_2 / n (N_1 + N_2). \quad (\text{Cf. RAO, 1952, p. 251.})$$

Mentionnons, à cet endroit, un travail de QUENOUILLE traitant du problème de l'élimination de variables insignifiantes en analyse discriminante (QUENOUILLE, 1947-49, p. 305).

1.53 *Le test de l'écart entre deux distances.*

Les tests précédents font intervenir les modifications apportées à la distance généralisée par l'inclusion ou l'exclusion d'une ou de plusieurs variables. Ils ne s'adressent pas directement à la différence $D_{p+q}^2 - D_p^2$ qu'implique le changement du nombre de variables. Aussi peut-on se demander s'il n'est pas possible de construire un test qui dépend explicitement de cette différence $D_{p+q}^2 - D_p^2$ ($q \geq 1$).

On peut montrer que lorsque les échantillons ont un grand effectif et qu'en même temps la valeur du paramètre Δ_p^2 , dans l'ensemble de base, est petite, la grandeur:

$$(241) \quad \frac{n - p + 1}{n + 1} \cdot \frac{n - p - q + 1}{q} \nu (D_{p+q}^2 - D_p^2), \quad \text{dans laquelle:}$$

$$n = N_1 + N_2 - 2, \quad \text{et:}$$

$$\nu = N_1 N_2 / n (N_1 + N_2),$$

suit approximativement une répartition $F(q; N_1 + N_2 - p - q - 1)$, (cf. RAO, 1952, p. 255).

Ce test a le défaut qu'il présuppose la connaissance de la valeur du paramètre Δ_p^2 , ainsi que celui d'être approximatif. En réalité, en effet, un test exact ferait ainsi intervenir ce paramètre en tant que « paramètre nuisible » (*nuiscent parameter* au sens de RAO) (cf. RAO, 1952, p. 201). Nous n'utiliserons pas ce test.

Au paragraphe 1.54, nous établirons une expression simple de cette différence $D_{p+q}^2 - D_p^2$, pour le cas où $q = 1$, susceptible d'une interprétation dans la terminologie de la régression linéaire multiple.

1.54 *La réduction de la distance par élimination d'une variable.*

En dépit de ce qu'il n'existe pas, à notre connaissance, de test exact concernant la différence, en distance, de deux discriminations utilisant un nombre inégal de variables, il vaut malgré tout la peine de chercher à interpréter cette différence. Nous nous limitons au cas particulier où, partant d'une discrimination à p variables, on élimine la dernière variable pour aboutir à une discrimination n'utilisant plus que $p - 1$ variables.

Pour examiner ce cas particulier, écrivons les équations déterminant les coefficients $h_{p,k}$ ($k = 1, \dots, p$) de la discrimination au moyen de p variables, sous la forme matricielle (cf. § 1.21):

$$(242) \quad [w_{kl}] [h_{p,k}] = [d_k] \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p.$$

La distance D_p^2 associée à cette discrimination sera:

$$(243) \quad D_p^2 = \sum_{k=1}^p d_k h_{p,k}.$$

Excluons la p -ième variable; les nouveaux coefficients seront $h_{p-1,k}$ ($k = 1, \dots, p-1$) et satisferont au système:

$$(244) \quad [w_{kl}] [h_{p-1,k}] = [d_k] \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p-1;$$

la distance associée à cette nouvelle discrimination sera:

$$(245) \quad D_{p-1}^2 = \sum_{k=1}^{p-1} d_k h_{p-1,k}.$$

Soit $[w_p^{kl}]$, $\left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p$, et $[w_{p-1}^{kl}]$, $\left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p-1$ les inverses des matrices $[w_{kl}]$ des systèmes 242 et 244 respectivement. La soustraction des $p-1$ premières équations correspondantes des systèmes 242 et 244 donne:

$$(246) \quad [w_{kl} (h_{p-1,k} - h_{p,k}) / h_{p,p}] = [w_{kp}], \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p-1.$$

Comparée aux $p-1$ premières équations du système déterminant les éléments w_p^{kp} ($k = 1, \dots, p$), à savoir:

$$(247) \quad [w_{kl}] [w_p^{kp}] = [\delta_{kp}] \quad \text{écrites sous la forme:}$$

$$(248) \quad [w_{kl} w_p^{kp} / (-w_p^{pp})] = [w_{kp}], \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p-1,$$

cette relation 246 implique:

$$(249) \quad (h_{p-1,k} - h_{p,k}) / h_{p,p} = w_p^{kp} / (-w_p^{pp}), \quad \text{donc:}$$

$$(250) \quad h_{p-1,k} - h_{p,k} = -h_{p,p} w_p^{kp} / w_p^{pp}, \quad k = 1, \dots, p-1;$$

δ est la notation de KRONECKER pour la fonction:

$$\delta_{kl} = \begin{cases} 1 & \text{lorsque } k = l \\ 0 & \text{lorsque } k \neq l. \end{cases}$$

On trouve de même pour relations entre w_p^{kl} et w_{p-1}^{kl} :

$$(251) \quad [w_{kl}] [w_p^{kl}] = [\delta_{kl}], \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p,$$

$$(252) \quad [w_{kl}] [w_{p-1}^{kl}] = [\delta_{kl}], \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p-1. \quad \text{En soustrayant:}$$

$$(253) \quad [w_{kl} (w_{p-1}^{kl} - w_p^{kl}) / w_{p,l}^p] = [w_{kp}], \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p-1.$$

La comparaison de cette relation avec l'équation 248 fournit l'égalité:

$$(254) \quad w_{p-1}^{kl} - w_p^{kl} = -w_p^{kp} w_p^{pl} / w_p^{pp}, \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p-1.$$

Les formules 250 et 254 permettent ainsi de réajuster les coefficients h d'un discriminant ainsi que les éléments w^{kl} de la matrice inverse lorsqu'on a éliminé une variable, sans qu'il soit nécessaire — et c'est là l'intérêt de ces formules — de refaire la résolution entière (cf. FISHER, 1958, § 29.1. Dans ses notations, w^{kl} est désigné par c). Des formules analogues peuvent être établies pour la situation inverse de l'inclusion d'une nouvelle variable (cf. COCHRAN, 1938, § 3).

Appliquons des résultats au calcul de la différence $D_p^2 - D_{p-1}^2$:

$$\begin{aligned}
 (255) \quad D_p^2 &= \sum_{k=1}^p h_{p,k} d_k \\
 &= \sum_{k=1}^p \{ h_{p-1,k} - (h_{p-1,k} - h_{p,k}) \} d_k \\
 &= \sum_{k=1}^p h_{p-1,k} d_k + \sum_{k=1}^p d_k \cdot h_{p,p} w_p^{kp} / w_p^{pp} .
 \end{aligned}$$

En convenant de poser $h_{p-1,k} = 0$, ce qui découle (formellement) de l'égalité 250 prolongée au cas où $k = p$, on peut remarquer que :

$$(256) \quad \sum_{k=1}^p h_{p-1,k} d_k = \sum_{k=1}^{p-1} h_{p-1,k} d_k = D_{p-1}^2 \quad \text{par définition.}$$

D'autre part :

$$\begin{aligned}
 (257) \quad \sum_{k=1}^p d_k \cdot h_{p,p} w_p^{kp} / w_p^{pp} &= (h_{p,p} / w_p^{pp}) \sum_{k=1}^p d_k w_p^{kp} \\
 &= (h_{p,p} / w_p^{pp}) \cdot h_{p,p} . \quad \text{Finalement :}
 \end{aligned}$$

$$(258) \quad D_p^2 = D_{p-1}^2 + (h_{p,p})^2 / w_p^{pp} , \quad \text{ou bien :}$$

$$(259) \quad D_p^2 - D_{p-1}^2 = (h_{p,p})^2 / w_p^{pp} .$$

Montrons encore que le terme $1/w_p^{pp}$ peut être interprété comme variabilité résiduelle dans une régression linéaire multiple. Plus exactement, considérons la régression $A_{ij(p)}$ de x_{ijp} en les variables $x_{ij1}, \dots, x_{ijp-1}$:

$$(260) \quad A_{ij(p)} = \bar{x}_p + \sum_{k=1}^{p-1} b_k (x_{ijk} - \bar{x}_k) \quad (k = 1, \dots, p-1) ;$$

les coefficients b_k satisfont aux équations :

$$(261) \quad [nw_{kl}] [b_k] = [nw_{kp}] , \quad n = N_1 + N_2 - 2 , \quad \text{ou bien :}$$

$$(262) \quad [w_{kl}] [b_k] = [w_{kp}] \left\{ \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p-1 .$$

La somme des carrés résiduelle nS_r (somme des carrés autour de la régression. Cf. § 1.12) sera donnée par :

$$\begin{aligned}
 (263) \quad nS_r &= \sum_{i,j} (A_{ij(p)} - x_{ijp})^2 \\
 &= nw_{pp} - \sum_{k=1}^{p-1} b_k (nw_{kp}) \\
 &= n \left\{ w_{pp} - \sum_{l=1}^{p-1} b_l w_{lp} \right\} .
 \end{aligned}$$

Les systèmes d'équations 244 et 262 ayant mêmes membres gauches, on peut écrire :

$$(264) \quad b_l = \sum_{k=1}^{p-1} w_{p-1}^{kl} w_{kp} , \quad (l = 1, \dots, p-1) , \quad \text{avec :}$$

$$(265) \quad w_{p-1}^{kl} = w_p^{kl} - w_p^{kp} w_p^{pl} / w_p^{pp} \quad \text{en vertu de l'équation 254.}$$

La symétrie de la matrice $[w_{kl}]$ permet d'écrire $w_{pl} = w_{lp}$, de sorte que :

$$\begin{aligned}
 (266) \quad w_{pp} - S_r &= \sum_{l=1}^{p-1} w_{pl} \left\{ \sum_{k=1}^{p-1} w_{kp} (w_p^{kl} - w_p^{kp} w_p^{pl} / w_p^{pp}) \right\} \\
 &= (1/w_p^{pp}) \left\{ w_p^{pp} \sum_{l=1}^{p-1} w_{pl} \sum_{k=1}^{p-1} w_{kp} w_p^{kl} - \sum_{l=1}^{p-1} w_{pl} w_p^{pl} \sum_{k=1}^{p-1} w_{kp} w_p^{kp} \right\} .
 \end{aligned}$$

Or, d'une part :

$$(267) \quad \sum_{k=1}^{p-1} w_{kp} w_p^{kl} = \left(\sum_{k=1}^p w_{kp} w_p^{kl} \right) - w_{pp} w_p^{pl} \\ = \delta_{pl} - w_{pp} w_p^{pl}, \quad \text{de sorte que:}$$

$$(268) \quad \sum_{l=1}^{p-1} w_{pl} \sum_{k=1}^{p-1} w_{kp} w_p^{kl} = \sum_{l=1}^{p-1} w_{pl} \delta_{pl} - w_{pp} \sum_{l=1}^{p-1} w_{pl} w_p^{pl} \\ = - w_{pp} \sum_{l=1}^{p-1} w_{pl} w_p^{pl}; \quad \text{d'autre part:}$$

$$(269) \quad \sum_{k=1}^{p-1} w_{kp} w_p^{kp} = \left(\sum_{k=1}^p w_{kp} w_p^{kp} \right) - w_{pp} w_p^{pp} \\ = \delta_{pp} - w_{pp} w_p^{pp} \\ = 1 - w_{pp} w_p^{pp}.$$

En substituant ces relations 268 et 269, on obtient :

$$(270) \quad w_p^{pp} (S_r - w_{pp}) = w_p^{pp} w_{pp} \sum_{l=1}^{p-1} w_{pl} w_p^{pl} + (1 - w_{pp} w_p^{pp}) \sum_{l=1}^{p-1} w_{pl} w_p^{pl} \\ = \sum_{l=1}^{p-1} w_{pl} w_p^{pl} \\ = 1 - w_{pp} w_p^{pp}, \quad \text{et finalement:}$$

$$(271) \quad w_p^{pp} = 1/S_r.$$

On conclut de ces considérations que l'élimination d'une variable, dans une discrimination, entraîne toujours une diminution de la distance :

$$(272) \quad D_p^2 - D_{p-1}^2 = (h_{p,p})^2 n S_r \geq 0,$$

diminution proportionnelle au carré du coefficient $h_{p,p}$ relatif à la variable x_p éliminée, et à la somme de carrés résiduelle dans la régression (linéaire multiple) de x_p en les autres variables. En d'autres termes, plus la dépendance (linéaire) entre x_p et les variables $x_1, \dots, x_{p-1}$ est forte, plus la diminution $D_p^2 - D_{p-1}^2$ sera petite. Cela n'a rien de surprenant : si la variable x_p présente de fortes corrélations avec les autres variables intervenant dans l'analyse discriminante, sa contribution à l'augmentation de la distance n'aura que peu d'importance.

Ecrite sous la forme :

$$(273) \quad D_{p-1}^2 = D_p^2 - (h_{p,p})^2 / w_p^{pp},$$

cette relation permet de déterminer directement — c'est-à-dire sans nouvelles résolutions — la distance D_{p-1}^2 de la discrimination à $p - 1$ variables, à partir des éléments de celle à p variables. Cette remarque pourrait revêtir un certain intérêt dans le problème du choix de la variable à exclure au cours d'un processus de simplification d'une discrimination comportant un grand nombre de variables.

1.6 LES TECHNIQUES ARITHMÉTIQUES

Il ressort des chapitres précédents qu'en ce qui concerne les opérations arithmétiques, deux difficultés risquent principalement d'apparaître du fait d'un nombre assez considérable de variables utilisées au départ, et de l'effectif relativement important d'individus observés dans l'investigation dont il sera plus spécialement question dans la seconde partie du présent travail. Ces difficultés concernent la résolution d'un système d'équations linéaires non-homogènes d'une part, l'inversion explicite d'une matrice d'autre part, c'est-à-dire la détermination de chaque élément de l'inverse. L'obstacle proprement dit est constitué par

le grand nombre de chiffres significatifs que, à moins que l'on se contente de résultats approximatifs, l'on est obligé de retenir dans les constantes de ces opérations en dépit de certaines transformations simplificatrices. Ce sont ces grands nombres qui rendent extrêmement fastidieuses, dans l'application pratique, des opérations élémentaires immédiatement réalisables en théorie. Il convient, dès lors, d'introduire une certaine planification des calculs garantissant l'exactitude et la précision des solutions tout en maintenant le volume du calcul numérique dans les limites acceptables pour la capacité usuelle d'une machine à calculer de pupitre. Ce sont ces schémas de calcul que nous voulons brièvement rappeler dans les trois prochains paragraphes.

1.61 *Résolution simple par éliminations successives.*

En présence d'un système de p équations linéaires non-homogènes en les inconnues h_k ($k = 1, \dots, p$) de la forme:

$$(274) \quad \sum_{k=1}^p h_k w_{kl} = w_{l,p+1} \quad (l = 1, \dots, p),$$

ou bien, sous forme matricielle:

$$(275) \quad [w_{kl}] [h_k] = [w_{l,p+1}] \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p,$$

on cherche une solution $h_1, \dots, h_p$ satisfaisant à ce système d'équations, ou bien un vecteur-colonne $[h_k]$, considéré comme matrice à 1 colonne et p lignes, vérifiant l'égalité matricielle ci-dessus. Pour unifier la notation, nous désignerons les seconds membres par $w_{l,p+1}$ ($l = 1, \dots, p$). La matrice $[w_{kl}]$ est carrée, d'ordre p et symétrique par hypothèse, c'est-à-dire:

$$(276) \quad w_{kl} = w_{lk} \quad \text{pour tout couple } (k, l).$$

Nous n'entendons pas, pour le moment, rechercher autre chose que la solution $[h_k]$. La méthode préconisée consiste alors à réduire le système d'ordre p à un système d'ordre $p - 1$ par simple élimination de l'une quelconque des inconnues, par exemple h_1 . Nous poserons:

$$(277) \quad w_{kl}^1 = w_{11} w_{kl} - w_{1k} w_{1l} \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p,$$

de sorte que:

$$(278) \quad w_{11}^1 = w_{11} w_{11} - w_{11} w_{11} = 0,$$

$$(279) \quad w_{k1}^1 = w_{11} w_{k1} - w_{1k} w_{11} = 0, \quad \text{et on aura un nouveau système:}$$

$$(280) \quad [w_{kl}^1] [h_k] = [w_{l,p+1}^1] \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 2, \dots, p \quad \text{d'ordre } p - 1.$$

En posant:

$$(281) \quad w_{kl}^2 = w_{22}^1 w_{kl}^1 - w_{2k}^1 w_{2l}^1,$$

on élimine h_2 de ce système, aboutissant ainsi à un système d'ordre $p - 2$:

$$(282) \quad [w_{kl}^2] [h_k] = [w_{l,p+1}^2] \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 3, \dots, p.$$

$p - 1$ telles éliminations successives conduisent enfin à l'équation:

$$(283) \quad [w_{pp}^{p-1}] [h_p] = [w_{p,p+1}^{p-1}], \quad \text{c'est-à-dire:}$$

$$(284) \quad w_{pp}^{p-1} h_p = w_{p,p+1}^{p-1}$$

qui détermine l'inconnue h_p . Celle-ci, introduite dans le système :

$$(285) \quad [w_{kl}^{p-2}] [h_k] = [w_{l,p+1}^{p-2}] \frac{k}{l} \Big\} = p - 1 \text{ ou } p ,$$

permettra de déterminer h_{p-1} , et ainsi de suite de proche en proche jusqu'à l'inconnue h_1 .

TABLEAU 11

k	Premier membre					Second membre	Contrôle
1	w_{11}	w_{12}	w_{13}	...	w_{1p}	$w_{1,p+1}$	$w_{1\tau}$
2	.	w_{22}	w_{23}	...	w_{2p}	$w_{2,p+1}$	$w_{2\tau}$
3	.	.	w_{33}	...	w_{3p}	$w_{3,p+1}$	$w_{3\tau}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
p	.	.	.	.	w_{pp}	$w_{p,p+1}$	$w_{p\tau}$
2	.	w_{22}^1	w_{23}^1	...	w_{2p}^1	$w_{2,p+1}^1$	$w_{2\tau}^1$
3	.	.	w_{33}^1	...	w_{3p}^1	$w_{3,p+1}^1$	$w_{3\tau}^1$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
p	.	.	.	.	w_{pp}^1	$w_{p,p+1}^1$	$w_{p\tau}^1$
1	1	2	3	...	p		

Pour le calcul pratique, il est commode de disposer les éléments selon le schéma indiqué au tableau 11; ce schéma ne reproduit que la première élimination. Il convient de remarquer que les éléments en dessous de la diagonale principale n'ont pas besoin d'être calculés du moment que les matrices $[w_{kl}]$, $[w_{kl}^1]$, etc., sont symétriques ainsi que l'indiquent l'hypothèse et la définition des grandeurs w_{kl}^1 , w_{kl}^2 , ..., w_{kl}^{p-1} . La dernière colonne du tableau sert de contrôle et fonctionne de la manière suivante. Par définition, on pose :

$$(286) \quad w_{l\tau} = \sum_{k=1}^{p+1} w_{kl} \quad (l = 1, \dots, p) .$$

Par analogie aux définitions ci-dessus, on calcule :

$$(287) \quad w_{l\tau}^1 = w_{11} w_{l\tau} - w_{1l} w_{1\tau} \quad (l = 2, \dots, p) ,$$

$$(288) \quad w_{l\tau}^2 = w_{22}^1 w_{l\tau}^1 - w_{2l}^1 w_{2\tau}^1 \quad (l = 3, \dots, p) , \text{ et ainsi de suite.}$$

Le contrôle consiste à vérifier, ligne par ligne, l'égalité :

$$(289) \quad w_{l\tau}^1 = \sum_{k=2}^{p+1} w_{kl}^1 \quad (l = 2, \dots, p) .$$

Car en effet, on a :

$$\begin{aligned}
 (290) \quad \sum_{k=2}^{p-1} w_{kl}^1 &= \sum_{k=2}^{p-1} (w_{11} w_{kl} - w_{1k} w_{11}) \\
 &= w_{11} \left(\sum_{k=2}^{p-1} w_{kl} \right) - w_{11} \left(\sum_{k=2}^{p-1} w_{1k} \right) \\
 &= w_{11} \left\{ \left(\sum_{k=1}^{p-1} w_{kl} \right) - w_{11} \right\} - w_{11} \left\{ \left(\sum_{k=1}^{p-1} w_{1k} \right) - w_{11} \right\} \\
 &= w_{11} w_{1\tau} - w_{11} w_{1\tau} = w_{1\tau}^1.
 \end{aligned}$$

En ce qui concerne la solution $[h_k]$, le contrôle le plus sûr consiste à vérifier, par substitution directe, que les valeurs h_k satisfont bien au système d'équations proposées.

Les grandeurs w_{11} , w_{22}^1 , w_{33}^2 , etc., jouent le rôle d'éléments-pivots, et la méthode qui vient d'être décrite est l'une des formes de la méthode dite de condensation pivotale (DWYER, 1951, p. 50). Pour conserver l'avantageuse propriété de symétrie d'un bout à l'autre de ce procédé d'élimination, il faut choisir les éléments-pivots sur la diagonale principale.

1.62 Détermination simultanée de la fonction discriminante et de la distance associée.

Un certain nombre de tests de signification sont basés non pas sur les valeurs h_k mêmes des coefficients, mais sur la distance D^2 , associée à la même discrimination et faisant intervenir ces coefficients implicitement seulement. On peut évidemment appliquer la définition :

$$D^2 = \sum_{k,l} w^{kl} d_k d_l,$$

mais lorsque, délibérément, on ne veut pas passer par les valeurs h_k , cette définition exige la connaissance de tous les éléments w^{kl} de la matrice inverse $[w_{kl}]^{-1}$. Or, l'inversion d'une matrice, en raison des interminables manipulations numériques qu'elle comporte, est l'un des procédés que l'on prend toutes les précautions d'éviter chaque fois que l'on peut. Il serait donc avantageux de disposer d'une méthode fournissant la valeur D^2 sans recours à l'inversion matricielle ni à la résolution complète des équations. La méthode que nous allons rappeler jouit précisément de cet avantage et permet en outre le calcul successif des valeurs D^2 pour un nombre croissant de variables incluses séparément l'une après l'autre. Cette méthode a ceci de remarquable, en plus, que par l'adjonction de certaines opérations complémentaires, mais nullement exigées pour le calcul des D^2 , la fonction discriminante peut être obtenue explicitement après l'inclusion de chaque nouvelle variable, s'il y a lieu.

Le tableau 12 reproduit le schéma de la méthode telle qu'elle a été indiquée par exemple par RAO en guise d'illustration numérique (RAO, 1952, p. 254). Nous en donnerons la justification algébrique dans le cas de $p = 4$ pour fixer les idées, et pour les trois premières phases seulement, la suite des opérations pouvant être déduite sans autre par analogie.

Soit donc l'égalité matricielle :

$$(291) \quad [w_{kl}][h_l] = [d_k]$$

définissant la discrimination en question. Le schéma ci-après (tabl. 12) repose essentiellement sur une forme particulière de la méthode de condensation pivotale, à savoir celle qui réduit la matrice $[w_{kl}]$ d'ordre p à une matrice $[w_{1,kl}]$ d'ordre $p - 1$ grâce à la transformation :

$$(292) \quad w_{1,kl} = w_{kl} - w_{k1} w_{1l} / w_{11},$$

l'élément w_{11} jouant le rôle d'élément-pivot ($w_{11} \neq 0$). On vérifie directement qu'en effet tous les éléments $w_{1,kl}$ s'annulent lorsque $k = 1$ ou lorsque $l = 1$.

TABLEAU 12

Dispersion				Ecart	Ligne
w_{11}	w_{12}	w_{13}	w_{14}	d_1	01
w_{21}	w_{22}	w_{23}	w_{24}	d_2	02
w_{31}	w_{32}	w_{33}	w_{34}	d_3	03
w_{41}	w_{42}	w_{43}	w_{44}	d_4	04
$w_{51} = d_1$	.	.	.	$d_5 = 0$	05
$w'_{11} = 1$	w'_{12}	w'_{13}	w'_{14}	d'_1	10
$w_{1.12} = w'_{12}$	$w_{1.22}$	$w_{1.23}$	$w_{1.24}$	$d_{1.2}$	11
$w_{1.31} = w'_{13}$	$w_{1.32}$	$w_{1.33}$	$w_{1.34}$	$d_{1.3}$	12
$w_{1.41} = w'_{14}$	$w_{1.42}$	$w_{1.43}$	$w_{1.44}$	$d_{1.4}$	13
$w_{1.51} = d'_1$	$w_{1.52} = d_{1.2}$	.	.	$d_{1.5}$	14
$w'_{1.21}$	$w'_{1.22} = 1$	$w'_{1.23}$	$w'_{1.24}$	$d'_{1.2}$	20
$w_{2.31}$	$w_{2.32} = w'_{1.23}$	$w_{2.33}$	$w_{2.34}$	$d_{2.3}$	21
$w_{2.41}$	$w_{2.42} = w'_{1.24}$	$w_{2.43}$	$w_{2.44}$	$d_{2.4}$	22
$w_{2.51}$	$w_{2.52} = d'_{1.2}$	$w_{2.53} = d_{2.3}$	.	$d_{2.5}$	23

La phase 0 (lignes 01 à 04) consiste à écrire explicitement les éléments w_{kl} de la matrice de dispersion, ainsi que les différences d_1 ; on y ajoute la ligne 05 où l'on définit:

$$(293) \quad w_{51} = d_1, \text{ et:}$$

$$(294) \quad d_5 = 0, \text{ les autres termes de la ligne restent momentanément indéterminés.}$$

La phase 1 débute par le calcul de la ligne pivotale 10, qui est la ligne 01 divisée par son premier élément (l'élément-pivot):

$$(295) \quad w'_{1l} = w_{1l}/w_{11} \quad (l = 1, \dots, 4),$$

$$(296) \quad d'_1 = d_1/w_{11}.$$

Les lignes 11, 12 et 13 contiennent la matrice réduite selon la transformation:

$$(297) \quad \left. \begin{aligned} w_{1.k} &= w_{kl} - w_{k1} w_{1l}/w_{11} \\ &= w_{kl} - w_{k1} \cdot w'_{1l} \end{aligned} \right\} k = 2, \dots, 4;$$

de même:

$$(298) \quad d_{1,k} = d_1 - w_{k1} d'_1 \quad (k = 1, \dots, 5); \quad \text{en particulier:}$$

$$(299) \quad d_{1,5} = d_5 - w_{51} d'_1.$$

La première ligne de cette matrice réduite est laissée de côté; quant aux éléments de la première colonne, on leur attribue, par définition nouvelle, la valeur des éléments de la ligne pivotale 10. En outre, on pose:

$$(300) \quad w_{1,51} = d'_1, \quad \text{et:}$$

$$(301) \quad w_{1,52} = d_{1,2} \quad \text{dans la ligne 14, les autres termes demeurent indéfinis.}$$

La matrice que forment les lignes 11, 12, 13 et 14 ainsi construites, est à son tour soumise à la même opération de condensation avec $w_{1,22}$ comme élément-pivot. Ce sera la phase 2 aboutissant, une fois les substitutions effectuées comme l'indique le schéma, aux lignes 21, 22 et 23. La ligne 20 en est la ligne pivotale obtenue en divisant la ligne 11 par l'élément $w_{1,22}$.

Le processus peut être ainsi prolongé, réduisant le nombre de lignes d'une unité lors de chaque phase, et il s'arrêtera de ce fait à la ligne 41 de la phase 4. Nous allons montrer maintenant que les lignes 14, 23 (32 et 41, non reproduites ici) contiennent tous les éléments numériques concernant la distance D_p^2 et le discriminant $H_p = \sum_{k=1}^p h_{p,k} x_k$ basés sur p variables; $p = 1, 2, 3, \dots$ successivement. Tout changement du nombre de variables, est-il besoin de le rappeler, entraîne, en général, une modification de chacun des coefficients du discriminant, d'où la nécessité de préciser par la notation $h_{p,k}$ à quel système de variables se rapporte ce coefficient. Il en est de même pour la distance D_p^2 .

Détermination des coefficients. — Dans le cas d'une seule variable, le coefficient $h_{1,1}$ satisfait à l'équation suivante:

$$(302) \quad w_{11} h_{1,1} = d_1, \quad \text{d'où immédiatement:}$$

$$(303) \quad h_{1,1} = d_1/w_{11} = w_{1,51};$$

c'est donc l'élément initial de la dernière ligne de la matrice obtenue dans la phase 1 (cf. tabl. 12).

Dans le cas de deux variables, les coefficients $h_{2,1}$ et $h_{2,2}$ satisfont aux équations:

$$(304) \quad w_{11} h_{2,1} + w_{12} h_{2,2} = d_1,$$

$$(305) \quad w_{21} h_{2,1} + w_{22} h_{2,2} = d_2.$$

L'élimination de $h_{2,1}$ conduit à l'équation:

$$(306) \quad (w_{22} - w_{12} w_{12}/w_{11}) h_{2,2} = d_2 - d_1 w_{12}/w_{11},$$

qui peut être écrite sous la forme:

$$(307) \quad w_{1,22} h_{2,2} = d_{1,2},$$

d'où il ressort que:

$$(308) \quad h_{2,2} = d_{1,2}/w_{1,22} = w_{2,52}. \quad \text{D'autre part:}$$

$$\begin{aligned} (309) \quad h_{2,1} &= d_1/w_{11} - w_{12} h_{2,2}/w_{11} \\ &= w_{1,51} - w_{1,21} d_{1,2}/w_{1,22} \\ &= w_{1,51} - w_{2,52} w'_{1,21} = w_{2,51}. \end{aligned}$$

Ainsi, les coefficients $h_{2,1}$ et $h_{2,2}$ sont, comme dans la phase précédente, les éléments initiaux de la dernière ligne de la matrice obtenue dans la phase 2.

En définitive, la phase p fournit les p coefficients $h_{p,k}$ ($k = 1, \dots, p$) du discriminant basé sur les p premières variables; ces coefficients sont les p éléments initiaux de la dernière ligne de la matrice relative à la phase p .

Détermination de la distance. — Dans le cas d'une seule variable (phase 1), on a par construction :

$$(310) \quad d_{1,5} = d_5 - w_{15} d'_1 = 0 - d_1 h_{1,1} = -D_1^2 \quad \text{par définition.}$$

La valeur de la distance, changée de signe, se trouve donc être le dernier élément de la ligne qui fournit, par ailleurs, le coefficient $h_{1,1}$.

Considérons maintenant la phase 2, c'est-à-dire le cas de deux variables. Par construction :

$$(311) \quad d_{2,5} = d_{1,5} - w_{1,52} d_{1,2}, \quad \text{ou bien, en revenant aux définitions :} \\ = d_1^2/w_{11} - d_{1,2}^2/w_{1,22}. \quad \text{Or :}$$

$$(312) \quad d_{1,2} = d_2 - w_{21} d_1/w_{11}, \quad \text{et :}$$

$$(313) \quad w_{1,22} = w_{22} - w_{21} w_{12}/w_{11}.$$

En substituant les expressions 312 et 313 dans l'égalité 311, $d_{2,5}$ prend la forme suivante :

$$(314) \quad d_{2,5} = -d_1^2/\{w_{11} - (w_{11} d_2 - w_{12} d_1)^2/w_{11} (w_{11} w_{22} - w_{12}^2)\},$$

ou encore, grâce aux dénominateurs communs :

$$(315) \quad d_{2,5} = -d_1 (d_1 w_{11} - d_2 w_{12})/(w_{11} w_{22} - w_{12}^2) - d_2 (d_2 w_{11} - d_1 w_{12})/(w_{11} w_{22} - w_{12}^2).$$

En regardant les équations définissant, dans le cas de deux variables, les coefficients $h_{2,1}$ et $h_{2,2}$, on reconnaît dans l'égalité 315 la valeur de ces solutions, de sorte que :

$$(316) \quad d_{2,5} = -\{d_1 h_{2,1} + d_2 h_{2,2}\} = -D_2^2 \quad \text{par définition.}$$

En d'autres termes, la valeur de la distance généralisée, changée de signe, apparaît à la fin de la ligne contenant également les coefficients $h_{2,1}$ et $h_{2,2}$.

La genèse des distances successives D_p^2 ainsi que des systèmes de coefficients $h_{p,1}, \dots, h_{p,p}$ ($p = 1, 2, 3, \dots$) est ainsi démontrée (cf. RAO, 1952, p. 254). En regardant de près le processus arithmétique tel qu'il est reproduit schématiquement au tableau 12, on constatera que les opérations concernant le calcul des coefficients h n'influencent d'aucune manière celles qui déterminent la distance. Lorsque seules les valeurs D_p^2 (pour différents p) sont requises, la partie inférieure gauche peut, par conséquent, être entièrement omise dans chaque phase. Il s'agit de la partie située en dessous de la diagonale principale des matrices, séparée, dans le schéma (tabl. 12) par le trait pointillé. On peut remarquer enfin que ce procédé conserve, en une certaine mesure, la symétrie des matrices, évitant ainsi l'obligation de calculer et d'inscrire deux fois certaines valeurs. Des simplifications d'écriture peuvent être ainsi introduites, de même que des totaux de contrôle à l'instar de ceux décrits au paragraphe précédent.

1.63 L'inversion d'une matrice.

Soit $[w_{kl}]$ ($k = 1, \dots, p$; $l = 1, \dots, p$) une matrice symétrique donnée. Son inverse $[w^{kl}] = [w_{kl}]^{-1}$ est définie par la relation :

$$(317) \quad [w_{kl}][w^{kl}] = [\delta_{kl}]$$

où δ_{kl} est le symbole de KRONECKER pour la fonction:

$$\delta_{kl} = \begin{cases} 0 & \text{si } k \neq l \\ 1 & \text{si } k = l \end{cases} .$$

En d'autres termes, $[\delta_{kl}]$ est la matrice unité. Ecrite explicitement, l'égalité 317 signifie:

$$(318) \quad \sum_{i=1}^p w_{ki} w^{il} = \delta_{kl} \quad \left. \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, p .$$

TABLEAU 13

Dispersion			Matrice unité			Contrôle	Ligne
w_{11}	w_{12}	w_{13}	1	0	0	$w_{1\tau}$	01
.	w_{22}	w_{23}	0	1	0	$w_{2\tau}$	02
.	.	w_{33}	0	0	1	$w_{3\tau}$	03
1	w'_{12}	w'_{13}	w'_{14}	0	0	$w'_{1\tau}$	10
.	$w_{1.22}$	$w_{1.23}$	$w_{1.24}$	1	0	$w_{1.2\tau}$	11
.	.	$w_{1.33}$	$w_{1.34}$	$w_{1.35}$	1	$w_{1.3\tau}$	12
.	1	$w'_{1.23}$	$w'_{1.24}$	$w'_{1.25}$	0	$w'_{1.2\tau}$	20
.	.	$w_{2.33}$	$w_{2.34}$	$w_{2.35}$	1	$w_{2.3\tau}$	21
.	.	1	$w'_{2.34}$	$w'_{2.35}$	$w'_{2.36}$	$w'_{2.3\tau}$	30
.	.	1	w^{13}	w^{23}	w^{33}	$w^{3\tau}$	33
.	1	.	w^{12}	w^{22}	w^{32}	$w^{2\tau}$	32
1	.	.	w^{11}	w^{21}	w^{31}	$w^{1\tau}$	31

Rechercher les éléments w^{kl} de la matrice inverse revient donc à résoudre les p systèmes non-homogènes d'équations à p inconnues tels qu'ils sont définis par les relations 318. Cette résolution se fera par exemple suivant la méthode de condensation pivotale décrite au paragraphe précédent (cf. également RAO, 1952, p. 31). En remarquant que la matrice $[w_{kl}]$ des coefficients des inconnues w^{il} est la même pour tous les systèmes, le schéma

de la résolution (si $p = 3$ pour fixer les idées) se mettra sous la forme ramassée indiquée au tableau 13, $w_{l\tau}$ étant la colonne du contrôle, à savoir :

$$(319) \quad w_{l\tau} = \sum_{k=1}^p w_{kl} + \delta_{ll} \quad (l = 1, \dots, p) .$$

Les lignes 10, 20 et 30 sont les lignes pivotales; les éléments-pivots sont w_{11} , $w_{1.22}$ et $w_{2.33}$. La ligne 33 est identique à la ligne 30. Les grandeurs-solutions w^{12} et w^{11} sont obtenues par substitution de w^{13} , donc :

$$(320) \quad w^{12} = w'_{1.24} - w'_{1.23} w^{13} \quad (\text{selon ligne 20}), \text{ et :}$$

$$(321) \quad w^{11} = w'_{14} - w'_{12} w^{12} - w'_{13} w^{13} \quad (\text{selon ligne 10}),$$

de même pour les autres colonnes. Les lignes 31, 32 et 33, prises dans cet ordre, contiennent ainsi les éléments de la matrice inverse.

Remarquons que :

$$(322) \quad \begin{aligned} w^{3\tau} &= w'_{1.2\tau} - w'_{1.23} w'_{2.3\tau} \\ &= (1 + w'_{1.23} + w'_{1.24} + w'_{1.25} + 0) - w'_{1.23} (1 + w'_{2.34} + w'_{2.35} + w'_{2.3\tau}) , \end{aligned}$$

donc également :

$$(323) \quad w^{3\tau} = 1 + w^{13} + w^{23} + w^{33} ,$$

toutes opérations effectuées. Cela signifie, comme l'indique le tableau 13 (coin inférieur gauche), qu'il faut inscrire une fois le nombre 1 sur chacune des lignes 33, 32 et 31 pour maintenir le bon fonctionnement de la colonne de contrôle.

Pour ne pas allonger cet exposé, nous avons laissé de côté la discussion de certaines transformations que l'on peut appliquer au départ aux systèmes d'équations dans le but d'une part de maintenir plus ou moins stable l'ordre de grandeur des coefficients intervenant, d'assurer l'exactitude arithmétique d'autre part (cf. WOOLF, 1951, p. 116).

2 APPLICATION

2.1 LES DONNÉES NUMÉRIQUES

2.11 *Délimitation de la liste primitive.*

Les individus sont des hommes ou des femmes de 20 à 49 ans dont on a relevé, entre autres, neuf caractères somatiques métriques au cours d'une enquête séro-anthropologique et génétique effectuée en 1954 par une équipe de travail dans diverses régions du canton des *Grisons (Suisse)* (cf. HAEGLER, KAUFMANN et LANG, 1959). Ces caractères, que nous appelons variables dans la partie 1, sont les suivants:

- (1) la taille assis (ou buste)
- (2) la longueur des jambes
- (3) le diamètre antéro-postérieur maximum du crâne
- (4) le diamètre transverse maximum du crâne
- (5) le diamètre bizygomatique du crâne
- (6) la hauteur faciale (morphologique)
- (7) la hauteur du nez
- (8) la largeur du nez
- (—) la stature (ou taille).

La longueur des jambes a été obtenue comme différence entre la stature et le buste; pour cette raison, la stature sera systématiquement écartée de toutes les discriminations et aucun numéro d'ordre ne lui a été assigné de ce fait.

Ces caractères sont numérotés comme il l'est indiqué ci-dessus, et le résultat brut de la mensuration sera désigné par X_{ijk} ($k = 1, \dots, 8$). Pour des raisons de commodité arithmétique, nous avons une fois pour toutes réduit toutes les mesures uniformément selon la transformation linéaire suivante:

$$(324) \quad x_{ijk} = X_{ijk} - Z_k,$$

et l'analyse numérique qui va suivre ne concernera que ces valeurs transformées exclusivement. La valeur des constantes Z_k est donnée au tableau 27 (§ 3.5), permettant ainsi de rétablir l'échelle originale s'il devait y avoir lieu. Par un abus de langage, nous conserverons à ces valeurs transformées leur désignation anthropologique dans l'échelle habituelle; nous les appellerons caractères tout court. Il devrait être évident que la transformation 324 n'affecte d'aucune manière les conclusions énoncées en fin du présent travail (cf. également § 1.13 au sujet des transformations linéaires).

Les individus considérés se répartissent, sans égard au sexe, en deux catégories définies par des critères linguistiques, ethnologiques, historiques et certainement aussi géographiques: les Walsers (abréviation W) et les Romanches (abréviation R). Nous désignerons par raciale toute spécification ayant trait à cette classification, sans vouloir attacher à ce terme aucune signification autre que celle d'une simple convention terminologique.

Chaque individu, walser ou romanche, peut en outre être caractérisé quant à son ascendance. C'est ainsi que par l'application de certains critères généalogiques, on distinguera les individus de type I de ceux du type II selon l'appartenance de leurs ascendants aux catégories walser et non-walser, ou bien romanche et non-romanche, le type II étant à cet égard considéré comme plus mélangé. Nous appellerons typique toute spécification se rapportant à cette subdivision I/II.

Nous obtenons ainsi les quatre catégories :

- WI, Walsers du type I (151 hommes, 126 femmes)
- WII, Walsers du type II (117 hommes, 134 femmes)
- RI, Romanches du type I (60 hommes, 54 femmes)
- RII, Romanches du type II (28 hommes, 10 femmes),

constituant, avec la différenciation sexuelle, la classification fondamentale de notre matériel numérique. Nous ne considérons, pour le moment, aucune subdivision plus fine, géographique par exemple. Ajoutons que seuls ont été retenus, pour l'analyse statistique, actuelle, les individus normaux dont on avait relevé chacun des huit caractères mentionnés sans exception.

Il convient de préciser très soigneusement que les Walsers dont il est question dans la liste primitive (§ 2.11) appartiennent tous à la catégorie qu'on a définie ailleurs comme orientale par opposition à la catégorie occidentale (cf. pour cela HOTZENKOECHERLE, 1944, p. 135). Plus précisément, ces Walsers orientaux comprennent des sujets mesurés à *Davos* et *Klosters* y compris les agglomérations attenantes, à *St. Antönien*, *Furna*, *Langwies* et *Wiesen*, à l'exclusion complète de *Tschappina* qui doit être rattachée à la catégorie occidentale. Les Romanches comprennent des habitants de *Salouf*, *Savognin* et *Tinizong* (vallée de l'*Oberhalbstein*).

Nous laissons le soin aux anthropologistes de justifier le choix de 20 à 49 ans comme âges limitant, selon eux, la classe des adultes; il se pourrait, en effet, qu'à 20 ans la croissance ne soit pas définitivement achevée pour tous les caractères ici considérés, qu'une erreur systématique s'introduise de ce fait. Il ne nous appartient pas davantage de faire la critique des critères établissant les subdivisions raciales et typiques bien que, de même en ce qui concerne la désignation des sujets qui auront été mesurés au cours de l'enquête, certains points ne nous paraissent pas absolument incontestables.

2.12 *Calculs préliminaires.*

Le matériel numérique ayant été ainsi délimité après exclusion des fiches étrangères ou incomplètes, le prochain pas a consisté à réarranger les données numériques sous une forme compacte qui les rend maniables pour le calcul. Une numérotation codifiée a été introduite contenant l'indication du domicile, du sexe, de l'appartenance raciale et typique ainsi qu'un numéro d'ordre pour chaque individu. Le changement d'échelle (eq. 324, § 2.11) a été effectué en même temps; de même, des colonnes de contrôle avec des totaux partiels ont été calculées en vue des vérifications. Pour chaque sujet, cette liste primitive indique ainsi, sur une ligne, la valeur (réduite) des huit caractères. Faute de place, nous renonçons à publier cette liste *in extenso*. Le tableau 27 (§ 3.5) par contre reproduit, outre l'effectif N , le total de chaque caractère pour les classes fondamentales et séparément pour les deux sexes, alors que le tableau 14 contient les moyennes correspondantes ainsi que la précision de ces moyennes sous la forme de la grandeur $s_k/\sqrt{N}$ (s_k = écart-type du k -ième caractère).

Il a fallu ensuite procéder au calcul des sommes de produits:

$$\sum_j x_{ijk} x_{ijl} \left\{ \begin{matrix} k \\ l \end{matrix} \right\} = 1, \dots, 8 \quad (p = 8),$$

dans les mêmes classes, contrôlées par la vérification des identités:

$$(325) \quad \sum_{k=1}^8 \left\{ \sum_j x_{ijk} x_{ijl} \right\} = \sum_j x_{ijl} x_{ij}^{\tau}, \quad \text{et:}$$

$$(326) \quad \sum_{l=1}^8 \left\{ \sum_j x_{ijl} x_{ij}^{\tau} \right\} = \sum_j (x_{ij}^{\tau})^2, \quad \text{avec, par définition:}$$

$$(327) \quad x_{ij}^{\tau} = \sum_{k=1}^8 x_{ijk},$$

les indices i et j se rapportant aux individus de l'une des classes considérées, τ désignant le total. Les tableaux 28 et 29 (§ 3.5) représentent ces sommes sous forme matricielle; les colonnes de contrôle ne seront plus utilisées par la suite et ont été laissées de côté.

TABLEAU 14

Moyennes $\bar{x}_k$ et précisions $s_k/\sqrt{N}$ de la moyenne des huit mesures dans les sous-populations principales ($k = 1, \dots, 8$; échelle transformée)

k	I ♂		II ♂		I ♀		II ♀	
	$\bar{x}_k$	$s_k/\sqrt{N}$	$\bar{x}_k$	$s_k/\sqrt{N}$	$\bar{x}_k$	$s_k/\sqrt{N}$	$\bar{x}_k$	$s_k/\sqrt{N}$
W a l s e r s (W)								
1	38.695	0.2668	38.701	0.2818	34.889	0.2440	35.425	0.2470
2	32.450	0.3309	31.855	0.3668	25.063	0.3528	25.634	0.3311
3	42.781	0.4768	41.248	0.5657	32.198	0.4926	32.582	0.5249
4	57.437	0.4449	55.684	0.4967	49.841	0.4836	49.731	0.4342
5	41.510	0.4106	40.385	0.5106	32.841	0.4561	33.022	0.3917
6	32.278	0.4705	31.359	0.5696	20.611	0.5066	22.082	0.4814
7	53.139	0.2829	53.718	0.3172	48.778	0.2750	49.067	0.2556
8	35.755	0.2117	35.419	0.2332	32.802	0.2344	32.672	0.2082
R o m a n c h e s (R)								
1	39.517	0.4175	39.929	0.6127	35.759	0.3812	33.000	1.1055
2	30.917	0.7214	31.643	1.0029	24.648	0.6041	24.500	1.1378
3	42.583	1.2484	40.179	1.2452	32.926	0.7670	32.800	1.2719
4	53.100	0.8264	53.821	1.2399	45.963	0.7373	46.300	1.5637
5	39.783	0.6605	40.357	1.2902	30.963	0.7098	31.400	1.0458
6	33.950	0.9871	32.607	1.2392	23.019	0.8962	18.100	1.6292
7	55.267	0.5019	53.357	0.6981	50.778	0.5377	49.800	1.1134
8	34.617	0.3172	35.821	0.5469	32.648	0.2874	32.000	0.8165

L'ensemble des tableaux 27, 28 et 29 (§ 3.5) permettent de reconstruire, en principe, toutes les constantes apparaissant dans les opérations subséquentes; ils contiennent en particulier tous les éléments nécessaires à la construction des systèmes d'équations déterminant les coefficients des diverses discriminations, ou bien des distances associées à ces discriminations. En ce qui concerne la reproduction, dans ce travail, des résultats numériques, nous n'avons intercalé des tableaux de chiffres (données, résultats intermédiaires ou finals) dans le corps du texte que pour autant qu'ils nous paraissaient faciliter la compréhension de celui-ci. Dans une annexe (le paragraphe 3.5) nous avons par conséquent réuni un certain nombre de tableaux de moindre nécessité, qui pourront pourtant être consultés à cet endroit chaque fois qu'il le faudra.

Ne disposant pas des facilités que nous aurait procurées le recours à une calculatrice électronique, nous n'avons pas reculé devant l'énorme tâche d'effectuer tous nos calculs

entièrement sur deux machines de pupitre électriques, une MADAS 20AZS principalement, et une FACIT ESA-O (à remise à zéro automatique). Une part importante du temps a ainsi dû être consacrée aux opérations préliminaires telles que le calcul des sommes de carrés et de produits.

2.13 *Problèmes anthropologiques et problèmes statistiques.*

Le matériel numérique ayant été ainsi définitivement mis au point, en particulier en ce qui concerne la localité de *Tschappina* (cf. § 2.11), il n'est peut-être pas inutile, avant de passer à la résolution des divers problèmes, de donner une vue d'ensemble de toute cette analyse numérique afin d'en montrer, d'une part, l'enchaînement des différentes parties, de faire ressortir d'autre part ce que chacun des deux points de vue, anthropologique ou statistique, comportait de spécifique.

D'emblée, l'enquête visait à l'étude des Walsers (orientaux). Anthropologiquement parlant, l'une des méthodes classiques pour la description d'une population consistant à en comparer les caractéristiques métriques ou descriptives à celles d'autres populations, on a aussi examiné des sujets non-walsers, à savoir des Romanches dans la vallée de l'*Oberhalbstein*.

L'élaboration des données métriques devait donc se faire en deux étapes : établir les caractéristiques générales pour les Walsers aussi bien que pour les Romanches, les comparer ensuite. Alors que la première étape, se réduisant à la simple détermination des paramètres de position (moyenne, médiane, mode) et de dispersion (le plus souvent l'étendue) des répartitions en question, est sans intérêt propre pour le statisticien parce que purement descriptive, la seconde devait par contre employer une technique spécifiquement statistique qui n'est guère en usage courant chez les anthropologistes : nous voulons parler des tests de signification. Cette seconde étape, celle des comparaisons, pouvait elle-même donner lieu à deux genres d'investigations, essentiellement : les comparaisons de populations caractère par caractère (chacun étant considéré indépendamment des autres) — méthode anthropologique par excellence — et les comparaisons basées sur l'emploi simultané de tous les caractères (métriques) relevés — méthode bien connue par les statisticiens sous le nom d'analyse à plusieurs dimensions. Dans l'un et l'autre cas, l'existence des tests préservent les conclusions des effets d'une interprétation subjective inhérente, en dépit de l'apparence d'une objectivité qu'assurerait l'usage des chiffres, à toute confrontation qualitative telle qu'elle est couramment pratiquée en anthropologie.

Le soin de comparer les Walsers aux Romanches quant aux caractères pris un à un, a été laissé aux anthropologistes, alors que le statisticien a préféré recourir aux méthodes multidimensionnelles dès le départ (cf. chap. 2.2). Quoi qu'il en soit, une des premières questions qui avait dû être résolue, était celle du type II. Considéré comme moins représentatif fallait-il le maintenir séparé du type I parce que trop différent de lui, ou au contraire, en vertu d'une similitude dûment établie, allait-on pouvoir dorénavant confondre les deux types, c'est-à-dire abandonner le concept même du type tel qu'il avait été défini ? Cette question exigeait le calcul des discriminations typiques (§ 2.21) effectuées, conformément à l'usage anthropologique, séparément pour les deux sexes ; ce calcul a conduit à l'exclusion, pour la suite de l'analyse discriminante, des sujets du type II.

Du côté anthropologique, l'étude du type II n'a pas été poursuivie spécialement ; du côté statistique, la position du type II par rapport au type I a été précisée par des tests adéquats (§ 2.51). L'examen des quatre discriminants typiques n'a pas été approfondi, mais il serait néanmoins intéressant de connaître leur comportement, en particulier pour savoir à quels caractères est due la différence entre le type I et le type II, pour autant que cette différence existe, c'est-à-dire est significative.

L'analyse s'est tout naturellement poursuivie par la discrimination raciale entre Walsers et Romanches, séparément pour les deux sexes toujours, et pour le type I seulement (§ 2.22). Elle a conduit à l'évidence de la différence raciale telle que les anthropologistes l'attendaient.

On pourrait être tenté de penser qu'à ces résultats — la mauvaise différenciabilité typique, la présence d'une dissemblance raciale très prononcée — joints aux tests concernant les écarts de moyennes des divers caractères dans les sous-populations envisagées, se limiterait l'apport statistique aux questions anthropométriques au sujet de la comparaison des Walsers et des Romanches. Cependant les statisticiens, une fois la réponse donnée aux questions des anthropologistes, ont à leur tour proposé et exécuté toute une série d'investigations concernant cette discrimination raciale, plus importante et plus intéressante dans notre enquête que la discrimination typique. Ces investigations avaient pour objet l'étude plus approfondie des divers caractères entrant dans la fonction discriminante, en vue d'une éventuelle simplification du procédé discriminatoire selon deux manières : premièrement (chap. 2.3) comparer les fonctions discriminantes masculine et féminine dans l'espoir de pouvoir en établir l'équivalence, ce qui conduirait à la construction d'une fonction commune valable pour les deux sexes indifféremment. Bien qu'artificielle aux yeux des anthropologistes conscients des différences sexuelles, si elle réussissait, cette construction aurait néanmoins la signification très précise suivante : les différences raciales entre Walsers et Romanches se manifestent de la même manière chez les hommes et chez les femmes. — Secondement (chap. 2.4) détecter ceux parmi tous les caractères relevés, qui jouent un rôle prépondérant dans la discrimination, pour les séparer de ceux qu'on pourrait négliger sans perte d'information préjudiciable (significative). On aurait alors un argument suffisant pour justifier la construction de fonctions ne dépendant plus que d'un nombre restreint de caractères. Quel que soit l'avantage résultant d'une telle réduction, la préoccupation de la valeur ou de l'importance d'un caractère relativement à un procédé comparatif déterminé, n'apparaît pas toujours très explicitement dans les travaux anthropologiques. Nous en voulons pour preuve, par exemple, que l'on voit la taille qui, selon la conception anthropologique même est un composé de différents segments dont le moins qu'on puisse dire est qu'ils ne présentent pas un mode de variation uniforme, continuer à jouir d'un prestige fondé vraisemblablement davantage sur la facilité avec laquelle cette mesure peut être prise, que sur sa vraie valeur, et ceci au détriment de caractères qui, tels le buste ou le tronc, sont maintenant reconnus lui être supérieur (cf. entre autres PIQUET, 1954, p. 57 et BONÉ, 1953, pp. 28-30).

C'est bien plus dans le but d'illustrer ce point de vue précisément que par souci de banale application numérique de méthodes relativement nouvelles, que nous n'avons pas hésité de pousser jusqu'à l'extrême ce procédé de simplification ainsi que la discussion des questions attenantes, en particulier celle de la corrélation et de l'ordre de succession (§ 2.43). Ainsi avons-nous consacré une très grande partie de nos efforts à la construction d'une fonction commune réduite aux trois ou quatre caractères principaux, et à la démonstration de l'équivalence de son pouvoir séparateur, quant à la discrimination raciale, à celui des deux fonctions discriminantes complètes. Cette fonction réduite commune, nous l'avons encore confrontée avec une fonction à coefficients entiers qui peut être considérée comme une approximation très grossière de la fonction exacte, établissant ainsi qu'une très grande précision dans la donnée des coefficients n'est peut-être pas absolument indispensable dans le cas particulier.

Enfin (chap. 2.5) ayant construit cette fonction réduite commune, nous l'avons exploitée dans l'analyse de diverses questions faisant intervenir les valeurs individuelles de cette fonction pour les sujets de cette enquête. La nécessité de préciser la situation du type II par exemple découlait naturellement de la discrimination typique. Profitant d'autre part de ce que nous avions calculé la valeur individuelle de la fonction pour chaque sujet de la liste primitive — entre autres —, nous n'avons pas estimé vain de regarder d'un peu plus près la répartition des valeurs individuelles dans les diverses sous-populations ici considérées, pour pouvoir préciser à cet égard certaines questions qui nous avaient été posées dans un autre domaine d'investigations à vrai dire (validité des tests usuels appliqués aux valeurs individuelles d'un discriminant, cf. § 2.53).

2.2 LES DISCRIMINATIONS FONDAMENTALES

2.21 *Les discriminations typiques.*

L'une des premières questions — et chronologiquement, et par ordre d'importance — posée par les anthropologistes était celle de l'attitude à observer à l'égard des individus du type II. Car si ceux-ci n'étaient pas réellement différents de ceux du type I, la distinction typique perdait son sens et n'avait plus de raison à être maintenue: les individus des deux types, aussi bien chez les Walsers que chez les Romanches, seraient dorénavant réunis en une seule catégorie appelée soit walsen tout court, soit romanche tout court. En respectant, selon l'usage anthropologique, la distinction sexuelle (désignée par les signes conventionnels ♀ et ♂), nous aurons donc les quatre discriminations typiques suivantes:

$$\begin{array}{ll} \text{WI } \sigma / \text{WII } \sigma & \text{RI } \sigma / \text{RII } \sigma \\ \text{WI } \varphi / \text{WII } \varphi & \text{RI } \varphi / \text{RII } \varphi . \end{array}$$

TABLEAU 15

Différences $d_k = \bar{x}_{1k} - \bar{x}_{2k}$ ($k = 1, \dots, 8$) des moyennes typiques dans les quatre sous-populations W ♂, W ♀, R ♂, R ♀, et test t relatif à cette différence (form. 51)

Discrimination	Caractères							
	1	2	3	4	5	6	7	8
WI♂/WII♂	—0.006 °	0.595 °	—1.533 *	1.753 *	1.125 °	0.919 °	—0.579 °	0.336 °
WI♀/WII♀	—0.536 °	—0.571 °	—0.384 °	0.110 °	—0.181 °	—1.471 *	—0.289 °	0.130 °
RI♂/RII♂	—0.412 °	—0.726 °	2.404 °	—0.721 °	—0.574 °	1.343 °	1.910 *	—1.204 *
RI♀/RII♀	2.759**	0.148 °	0.126 °	—0.337 °	—0.437 °	4.919 *	0.978 °	0.648 °
° différence non significative * différence significative ($P = 0.05$) ** différence hautement significative ($P = 0.01$)								

Le tableau 15 indique, pour les huit caractères et dans les catégories envisagées, les différences de moyennes constituant les membres droits des équations en h_k ; il donne également le résultat du test de la différence de ces deux moyennes, effectué selon la formule 51. Les matrices de dispersion sont reproduites au tableau 30 (§ 3.5) et les solutions h_k au tableau 16. La résolution des équations:

$$[w_{kl}] [h_k] = [d_l] \quad (\text{cf. eq. 92}),$$

a été effectuée sur le modèle du tableau 11 (§ 1.61); ordre d'élimination des inconnues: 1-2-3-4-5-6-7-8. Le tableau 16 contient encore les distances D_8^2 et le résultat du test de cette distance selon la formule 134, avec $f_1 = 8$ et $f_2 = N - 9$. Ces tests montrent que la discrimination entre le type I et le type II n'est significative, au seuil de signification de 5%, que chez les Walsers masculins. Cela fait que nous maintenons, pour le moment, la notion du type, mais que nous écartons les individus du type II de toutes les discriminations qui vont suivre.

TABLEAU 16

Coefficients h_k ($k = 1, \dots, 8$), distances D_8^2 et test F de cette distance (form. 134) pour les quatre discriminations typiques

	Discrimination			
	WI ♂/WII ♂	WI ♀/WII ♀	RI ♂/RII ♂	RI ♀/RII ♀
h_1	— 0.0584	— 0.0552	— 0.0754	0.3669
h_2	0.0376	— 0.0218	— 0.0440	— 0.0857
h_3	0.0246	0.0090	0.0392	— 0.1342
h_4	0.0575	0.0101	— 0.0408	0.0046
h_5	— 0.0049	— 0.0012	0.0598	— 0.1343
h_6	0.0487	— 0.0408	— 0.0181	0.1811
h_7	— 0.1049	0.0057	0.1529	— 0.0371
h_8	0.0248	0.0116	— 0.2114	0.0993
D_8^2	0.2696	0.0998	0.6746	1.9587
F	2.163 *	0.788 °	1.479 °	1.833 °
f_1	8	8	8	8
f_2	259	251	79	55

° valeur non significative.
* valeur significative ($P = 0.05$).

2.22 Les discriminations raciales complètes.

Nous désignons par discriminant racial complet l'une ou l'autre des deux fonctions des huit caractères ($p = 8$) établies en vue de comparer les Walsers I aux Romanches I, séparément chez les deux sexes; ce seront donc les discriminations WI ♂/RI ♂ et WI ♀/RI ♀ respectivement. La moyenne des caractères et leur différence figurent au tableau 17 avec

TABLEAU 17

Différence $d_k = x_{1k} - x_{2k}$ ($k = 1, \dots, 8$) des moyennes raciales dans les deux sous-populations I ♂, I ♀, et test t relatif à cette distance (form. 51)

Diseri- mination	Caractères							
	1	2	3	4	5	6	7	8
WI ♂/RI ♂	— 0.822 °	1.533 *	0.198 °	4.337 **	1.727 °	— 1.672 °	— 2.128 **	1.138 **
WI ♀/RI ♀	— 0.870 °	0.415 °	— 0.728 °	3.878 **	1.878 °	— 2.408 °	— 2.000 **	0.154 °

° différence non significative.
* différence significative ($P = 0.05$).
** différence hautement significative ($P = 0.01$).

le résultat du test de cette différence (test de t , eq. 51), les matrices de dispersion $[w_{kl}]$ au tableau 31 (§ 3.5), et les solutions h_k ainsi que la distance D_8^2 avec le résultat du test (selon eq. 134) au tableau 18. La résolution des équations:

$$[w_{kl}][h_k] = [d_k] \quad (\text{cf. eq. 92}),$$

a été effectuée selon le schéma du tableau 12 (§ 1.62), l'ordre d'élimination étant l'ordre naturel 1-2-3-4-5-6-7-8. Les deux rapports $F(f_1; f_2)$, où $f_1 = 8$ et $f_2 = N - 9$, sont significatifs à 0.1 %, indiquant par là une différence très marquée, chez les individus du type I, entre Walsers et Romanches indépendamment du sexe et compte tenu de l'ensemble des huit caractères.

TABLEAU 18

Coefficients h_k ($k = 1, \dots, 8$), distances D_8^2 et test F de cette distance (form. 134) pour les deux discriminations raciales complètes

	Discrimination		
	WI ♂/RI ♂	WI ♀/RI ♀	WI (♂; ♀)/RI (♂; ♀)
h_1	— 0.2079	— 0.1250	— 0.1633
h_2	0.1463	0.0811	0.1083
h_3	— 0.0146	— 0.0266	— 0.0217
h_4	0.1638	0.1785	0.1670
h_5	— 0.0502	0.0062	— 0.0288
h_6	0.0120	— 0.0352	— 0.0123
h_7	— 0.1954	— 0.1790	— 0.1871
h_8	0.1531	— 0.0093	0.0778
D_8^2	1.5866	1.3071	1.3570
F	8.230 **	5.933 **	13.453 **
f_1	8	8	8
f_2	202	171	382
** valeur hautement significative ($P = 0.01$).			

Ceci n'a, du reste, rien de surprenant puisque tant chez les hommes que chez les femmes plusieurs caractères montrent des écarts significatifs même lorsqu'ils sont considérés seuls, ainsi que le montre l'application du test de t selon la formule 51 dont les résultats figurent dans les tableaux 15 (§ 2.21) et 17. Ces résultats, à leur tour, suggèrent immédiatement la question de l'importance des différents caractères dans l'une ou l'autre des fonctions discriminantes, à supposer que cette question n'ait pas été prévue dès le début de l'analyse statistique. Le chapitre 2.4 sera précisément consacré à des considérations de ce genre.

Aux figures 1 et 2, nous avons tenté une représentation géométrique de la disposition spatiale des sous-populations walser et romanche, séparément pour les deux sexes. Pour des raisons qui s'éclairciront par la suite nous avons choisi les trois caractères buste (1), diamètre transverse (4) et hauteur du nez (7), et reporté chaque groupe ($x_1; x_4; x_7$) de valeurs observées comme point de coordonnées x_1, x_4, x_7 , dans un espace à trois dimensions repéré par un système trirectangle d'axes cartésiens dont les figures 1 et 2 seraient les images

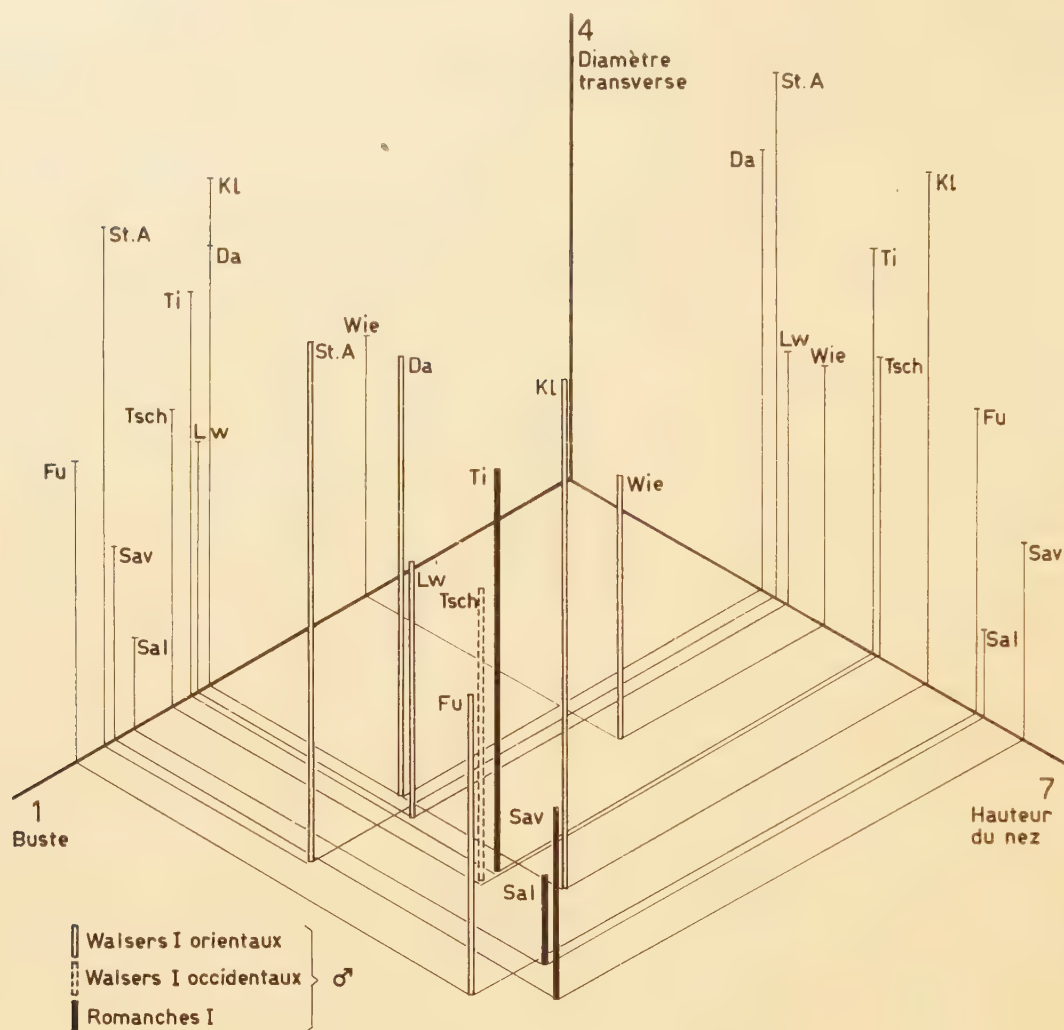


FIG. 1.

Groupement des sujets masculins du type I selon les localités, en fonction des moyennes des trois caractères buste, diamètre transverse et hauteur du nez.

perspectives isométriques (cf. STIEFEL, 1947, p. 41). Ce que nous avons représenté, en réalité, ce sont les points représentatifs des moyennes $\bar{x}_1$, $\bar{x}_4$, $\bar{x}_7$ des six localités walsers orientales (voir § 2.11), d'une localité walser occidentale (*Tschappina*) et des trois localités romanches de l'*Oberhalbstein*, points situés au sommet de la verticale correspondante (trait renforcé). Ont également été représentées les projections orthogonales de la configuration de ces points sur les faces du trièdre de référence (trait fin). Abstraction faite de *Tschappina* (Tsch) et sans vouloir entrer dans le détail de cette représentation, on remarque cependant le groupement assez net des localités walsers d'une part, celui des romanches, d'autre part) avec des exceptions comme *Furna* (Fu), et *Tinizong* (Ti) chez les hommes, *St. Antönien* (St.A, chez les femmes, exceptions qui continueront à se manifester de manière assez constante tout au long de la discrimination. Nous laissons aux anthropologistes le soin d'expliquer ces exceptions.

2.3 COMPARAISONS

2.31 Comparaisons des deux discriminations raciales complètes.

L'examen attentif des coefficients h_k relatifs aux deux discriminants raciaux complets (cf. tabl. 18, § 2.22), ou mieux encore celui des valeurs $h_k/\sqrt{w_{kk}}$ qui est une estimation de l'importance (relative) du coefficient h_k , révèle (fig. 3) une certaine similitude de structure

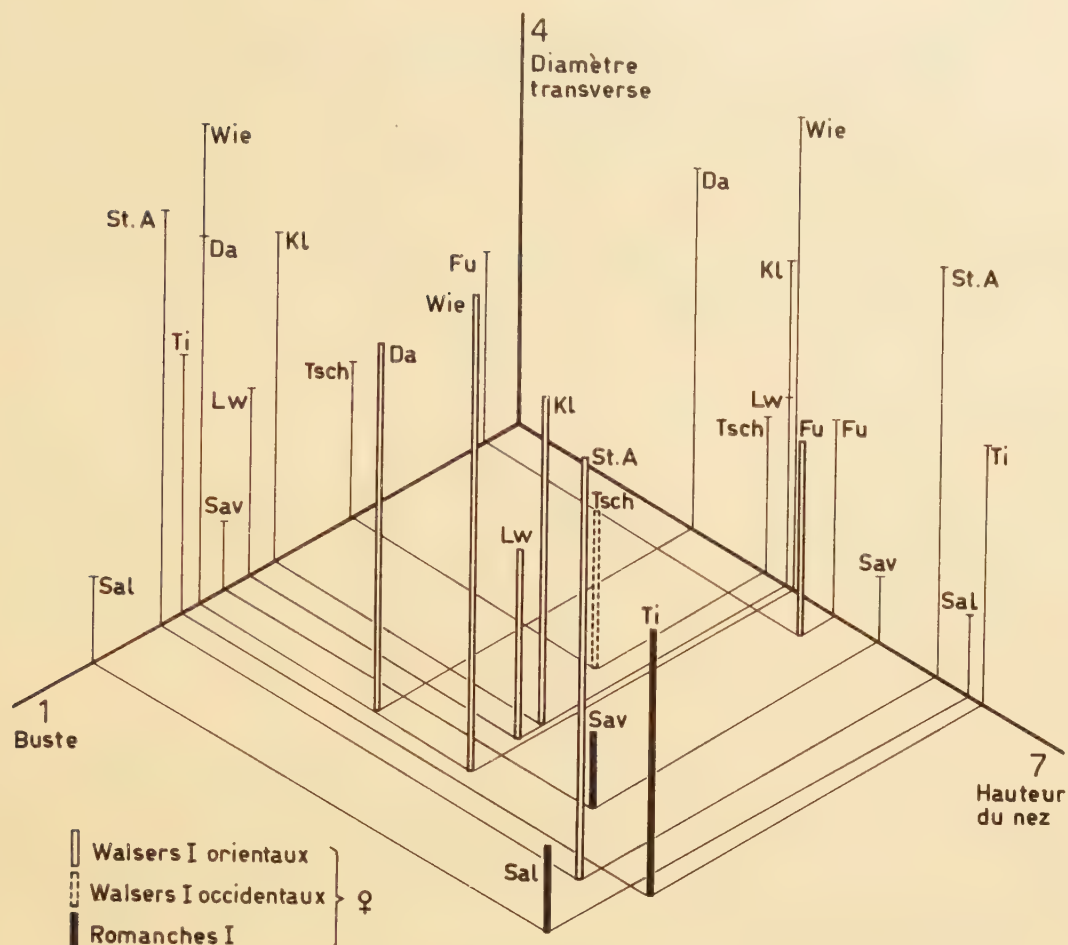


FIG. 2.

Groupement des sujets féminins du type I selon les localités, en fonction des moyennes des trois caractères buste, diamètre transverse et hauteur du nez.

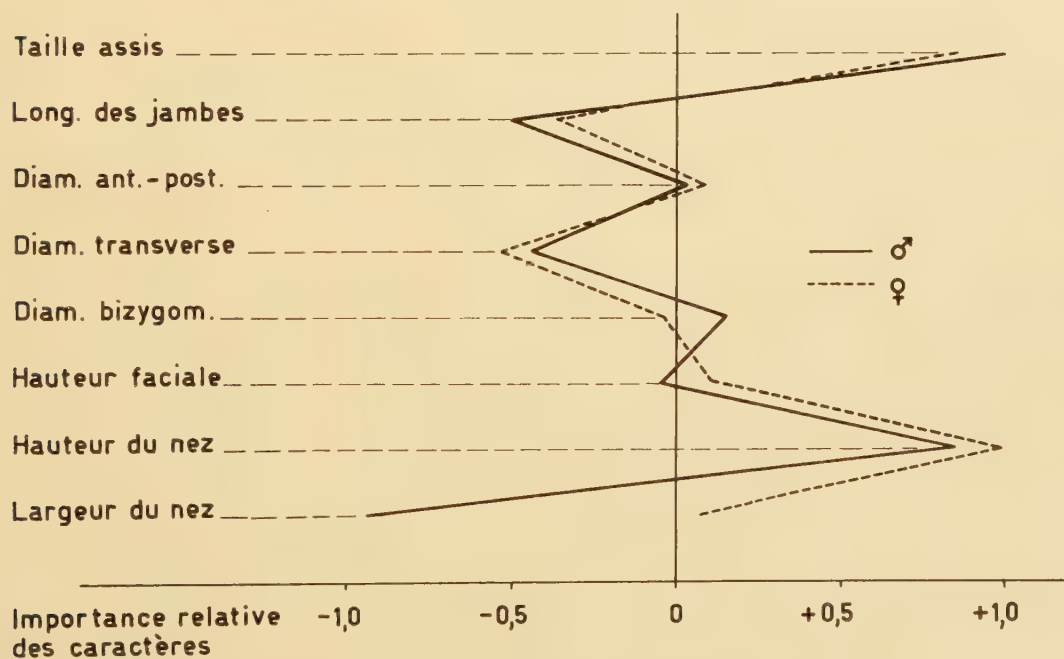


FIG. 3.

Importance des divers caractères intervenant dans les discriminations raciales complètes (échelle normée).

des deux fonctions en question. Quelque grossière que puisse être cette estimation, elle incite néanmoins à penser que les deux discriminants pourraient être équivalents. Nous allons préciser cette question au moyen du test :

$$\frac{(n - p - 1) \psi}{np} \sum_{k,l} w^{kl} (d'_k \pm d''_k) (d'_l \pm d''_l) = F(f_1; f_2) \quad (\text{cf. § 1.43}).$$

Les quatre échantillons sont ici les quatre sous-populations WI ♂, WI ♀, RI ♂, et RI ♀ dont les effectifs N_i ($i = 1, \dots, 4$) sont donnés au tableau 27 (§ 3.5), et les différences d'_k, d''_k au tableau 17 (§ 2.22). La matrice de dispersion $[w_{kl}]$ définie par :

$$nw_{kl} = \sum_{i=1}^4 \sum_{j=1}^{N_i} (x_{ijk} - \bar{x}_{ik}) (x_{ijl} - \bar{x}_{il}), \text{ avec :}$$

$$n = N_1 + N_2 + N_3 + N_4 - 4 ,$$

peut être calculée d'après les données des tableaux 27, 28 et 29 (§ 3.5); son inverse $[w^{kl}]$ a été déterminée selon le schéma du tableau 13 (§ 1.63). Ces deux matrices constituent le tableau 32 (§ 3.5).

Avec $n = 387$, $p = 8$, $1/\psi = \sum_i (1/N_i) = 0.049\,746$, on trouvera pour valeurs F :

$$\left. \begin{array}{l} \text{avec signes } + : F(f_1; f_2) = 13.318^* \\ \text{avec signes } - : F(f_1; f_2) = 0.790^o \end{array} \right\} f_1 = 8; f_2 = 380 .$$

Le calcul des valeurs F s'est effectué en deux étapes : le calcul des éléments de la matrice $[(d'_k + d''_k) (d'_l + d''_l)]$ ou $[(d'_k - d''_k) (d'_l - d''_l)]$, où $k = 1, \dots, 8$ et $l = 1, \dots, p$ (cf. tabl. 33, § 3.5), suivi de la multiplication accumulative de ces éléments par $[w^{kl}]$, terme par terme. La seconde de ces valeurs F n'étant pas significative, on en déduit l'équivalence des deux fonctions discriminantes considérées. En d'autres termes, les données observées ne sont pas en contradiction avec l'hypothèse :

$$E(d'_k) = E(d''_k) , \quad k = 1, \dots, 8 .$$

Pour l'anthropologiste, cela veut dire qu'en moyenne et pour l'ensemble des caractères, la différence entre Walsers et Romanches est la même chez les deux sexes pour autant qu'il s'agit du type I seulement.

Rien ne nous empêche dès lors de réunir, chez les Walsers aussi bien que chez les Romanches, hommes et femmes en une seule catégorie et de recalculer un discriminant racial complet commun pour cette nouvelle classification WI (♂, ♀)/RI (♂, ♀). La fonction ainsi obtenue (cf. tabl. 18, § 2.22) servirait à la discrimination raciale sans égard au sexe. Les éléments w_{kl} et d_l des équations :

$$[w_{kl}] [h_k] = [d_l] ,$$

calculés d'après les données des tableaux 27, 28 et 29 (§ 3.5), figurent au tableau 34 (§ 3.5). Il est évident que l'existence d'une telle fonction commune réalise une simplification considérable, formellement du moins, du moment qu'un seul discriminant peut être appliqué indifféremment dans toutes les situations dans lesquelles il fallait recourir, auparavant, à l'une ou l'autre de deux formules différentes suivant le sexe. Cette simplification toutefois, si défendable fût-elle, pourrait ne pas rester incontestée. Du point de vue statistique déjà, on pourrait en effet objecter que la réunion des deux classes masculine et féminine, en augmentant la dispersion des caractères, ferait baisser le pouvoir discriminatoire de la fonction commune. Car on sait bien que pour la plupart des caractères anthropométriques, il existe des différences dues au sexe. Dans une discrimination commune, c'est-à-dire celle où le sexe n'est pas considéré, ces différences pourraient donner lieu à des interactions avec les différences raciales et gêner par là la discrimination : il serait difficile de dissocier la différence sexuelle de la différence raciale. Quant au point de vue du praticien anthropologiste, le discriminant commun pourrait paraître quelque peu artificiel étant donné que la tendance

moderne est à l'étude séparée des deux sexes tout en n'excluant pas pour autant, les comparaisons entre hommes et femmes, cela va sans dire. On peut alors remarquer que l'analyse présente, comme du reste toute cette enquête plus généralement, ne se propose pas de mettre en évidence des différences sexuelles depuis longtemps connues. Autrement dit, il n'y a pas d'inconvénient à étudier hommes et femmes en bloc pour mettre en évidence des différences raciales (ou typiques).

Une étude plus détaillée, c'est-à-dire respectant à la fois la différenciation sexuelle et une échelle chronologique (classes d'âge) a été faite sur le matériel walser par LINDER, mais cependant pour le seul caractère buste uniquement (LINDER, 1960, § 515.2).

2.32 Comparaisons basées sur les discriminations typiques.

Considérons les deux discriminations relatives à des sujets masculins et féminins, respectivement, telles que les déterminent les coefficients g_k du tableau 19. En spécifiant qu'il s'agit de deux discriminants typiques, on pourrait songer à les comparer aux deux couples de discriminants typiques du tableau 16 (§ 2.21). On ne manquera pas de constater la similitude des fonctions proposées ci-dessous avec les fonctions correspondantes, quant au sexe, de la discrimination typique chez les Walsers.

TABLEAU 19

Coefficients g_k ($k = 1, \dots, 8$) de deux discriminations typiques antérieures

Discrimination	Coefficients							
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
♂	—0.0629	0.0490	0.0308	0.0551	—0.0069	0.0404	—0.0895	0.0365
♀	—0.0629	—0.0170	0.0028	0.0133	—0.0152	—0.0448	0.0053	0.0049

Appliquons le test de comparaison (§ 1.42) en supposant pour le moment que les coefficients g_k (tabl. 19) sont ceux de deux discriminants $G = \sum_k g_k x_k$ désignés par avance; nous approfondirons cette supposition par la suite. Ce test, selon la formule 215, exige la connaissance de la distance D_G^2 associée à la fonction G :

$$D_G^2 = (\bar{G}_1 - \bar{G}_2)^2 / \text{var}(G) \quad (\text{selon eq. 206}), \text{ avec:}$$

$$\bar{G}_1 - \bar{G}_2 = \sum_k g_k d_k, \quad \text{et:}$$

$$\text{var}(G) = \sum_{k,l} g_k g_l w_{kl} \quad (\text{selon eq. 207}).$$

En partant des différences d_k du tableau 15 (§ 2.21) et des coefficients g_k du tableau 19, on trouve:

$$\bar{G}_1 - \bar{G}_2 = 0.2667 \quad (\text{♂}),$$

$$\bar{G}_1 - \bar{G}_2 = 0.1115 \quad (\text{♀}),$$

tandis que les quantités w_{kl} du tableau 30 (§ 3.5), pour la discrimination walser, donnent en fin de compte:

$$\text{var}(G) = 0.2717 \quad (\text{♂}),$$

$$\text{var}(G) = 0.1342 \quad (\text{♀}).$$

Avec ces valeurs, nous obtenons pour distances D_G^2 :

$$D_G^2 = 0.2717 \quad (\mathfrak{J}) ,$$

$$D_G^2 = 0.0927 \quad (\mathfrak{Q}) ,$$

alors que les distances associées aux discriminations typiques H valent, pour les Walsers (tabl. 16, § 2.21):

$$D_H^2 = 0.2696 \quad (\mathfrak{J}) ,$$

$$D_H^2 = 0.0998 \quad (\mathfrak{Q}) .$$

On a en outre les constantes suivantes (selon tabl. 27, § 3.5):

$$N_1 = 151, \quad N_2 = 117, \quad N = N_1 + N_2 = 268, \quad n = N - 2 = 266, \quad \nu = 0.2478 \quad (\mathfrak{J}) ;$$

$$N_1 = 126, \quad N_2 = 134, \quad N = N_1 + N_2 = 260, \quad n = N - 2 = 258, \quad \nu = 0.2517 \quad (\mathfrak{Q}) .$$

De cette façon, on trouve pour valeur U selon la formule 213:

$$U = \frac{1 + \nu D_H^2}{1 + \nu D_G^2} - 1 ,$$

$$U = 0.001842 \quad (\mathfrak{J}) , \quad \text{et} \quad U = 0.001748 \quad (\mathfrak{Q}) ,$$

et finalement (formule 214):

$$F(f_1; f_2) = (N - p - 1) U / (p - 1) ,$$

avec $f_1 = p - 1$ et $f_2 = N - p - 1$; $p = 8$. Ainsi:

$$F(7; 259) = 0.068^\circ \quad (\mathfrak{J}) ; \quad F(7; 251) = 0.063^\circ \quad (\mathfrak{Q}) .$$

Ces valeurs F n'étant significative ni l'une ni l'autre, on en conclut que chez les hommes comme chez les femmes, les fonctions H et G peuvent être regardées comme équivalentes. Plus exactement, la distance D_H^2 associée au meilleur discriminant (linéaire) H déterminé par les données observées, ne doit pas être considérée comme significativement supérieure à la distance D_G^2 de la fonction proposée G ; cette conclusion vaut pour les deux sexes.

Il convient de préciser, maintenant, que les coefficients g_k du tableau 19 sont ceux de deux discriminants typiques calculés à partir d'un matériel numérique légèrement différents en ce qu'il comportait, outre les données précédemment délimitées concernant les Walsers orientaux exclusivement (cf. § 2.11), un certain nombre d'observations se rapportant à des Walsers occidentaux. Il s'agit de la localité de *Tschappina* incluse dans l'enquête, mais reconnue comme occidentale au cours des travaux de dépouillement seulement, et devant par conséquent être exclue de l'analyse présente qui, par principe, ne concerne que les Walsers orientaux. A la suite de ce remaniement des données, il a fallu recalculer entièrement certaines discriminations précédemment établies en tenant compte des données de *Tschappina*. Disposant des deux séries de discriminants, il était naturel de chercher à les comparer afin de faire ressortir l'effet — non significatif dans le cas particulier — de l'exclusion en question. Ce remaniement expliquera aussi certaines petites différences dans les résultats numériques du présent travail par rapport à des travaux antérieurs.

2.4 SIMPLIFICATION

2.41 Les caractères importants pour la discrimination raciale.

L'application du test t (form. 51) montre que toutes les moyennes de caractère $\bar{x}_k$, chez les Walsers, ne s'écartant pas significativement de la moyenne correspondante romanche; et cette affirmation vaut aussi bien pour les hommes que pour les femmes (cf. tabl. 17,

§ 2.22). Tous ces caractères pourraient, par conséquent, ne pas avoir une égale importance dans le discriminant, ce qui se traduit déjà dans le comportement de la grandeur $h_k/\sqrt{\bar{w}_{kk}}$ (fig. 3, § 2.31). Il s'agit donc de détecter ceux parmi les huit caractères qui exercent une influence primordiale dans la discrimination, pour les distinguer de ceux qui sont secondaires.

En respectant l'ordre de succession des caractères défini par l'indice k (cf. § 2.11, ou tabl. 27, § 3.5), la valeur D_1^2 basée sur l'unique premier caractère a été déterminée, puis les valeurs D_p^2 ($p = 2, 3, 4, \dots, 8$) en adjoignant un à un les caractères dans l'ordre précité. Le schéma du tableau 12 (§ 1.62) a été utilisé sous sa forme simplifiée, la connaissance des fonctions H_p successives n'étant pas exigée pour le moment.

TABLEAU 20

Distances D_p^2 , accroissements $D_p^2 - D_{p-1}^2$ et test F relatif à cet accroissement (form. 240) pour deux discriminations raciales; ordre d'inclusion: 1-2-3-4-5-6-7-8

Caractères inclus	p	Discrimination ♂			Discrimination ♀		
		D_p^2	$D_p^2 - D_{p-1}^2$	F	D_p^2	$D_p^2 - D_{p-1}^2$	F
1	1	0.0597	...	2.563 °	0.0947	...	3.580 °
12	2	0.3326	0.2729	11.521 **	0.1087	0.0140	0.516 °
123	3	0.3331	0.0005	0.020 °	0.1098	0.0011	0.040 °
1234	4	0.9941	0.6610	26.183 **	0.7801	0.6703	24.343 **
12345	5	1.0122	0.0181	0.633 °	0.7849	0.0048	0.152 °
123456	6	1.1304	0.1182	4.101 *	1.0229	0.2380	7.495 **
1234567	7	1.4395	0.3091	10.462 **	1.3067	0.2838	8.516 **
12345678	8	1.5866	0.1471	4.715 *	1.3071	0.0004	0.014 °
° valeur non significative. * valeur significative ($P = 0.05$). ** valeur hautement significative ($P = 0.01$).							

Le tableau 20 reproduit les valeurs D_p^2 successives relativement aux deux discriminations WI ♂/RI ♂, et WI ♀/RI ♀, ainsi que les accroissements $D_p^2 - D_{p-1}^2$. Le test de D_1^2 a été appliqué selon la formule 134, tandis que les D_p^2 ont été testés par rapport aux D_{p-1}^2 selon la formule 240; le résultat de ces tests figure au tableau 20 sous la forme de l'astérisque conventionnel. Les valeurs N_i , N , n , ν et η utiles pour les tests sont les suivantes:

$$N_1 = 151, N_2 = 60, N = 211, n = 209, \eta = 42.938\ 389, \nu = 0.205\ 447 \quad (\text{♂});$$

$$N_1 = 126, N_2 = 54, N = 180, n = 178, \eta = 37.800\ 000, \nu = 0.212\ 360 \quad (\text{♀});$$

(cf. tabl. 27, § 3.5).

De ces tests on peut tirer diverses conclusions: tant chez les hommes que chez les femmes, l'introduction des caractères 4 (diamètre transverse), 6 (hauteur faciale) et 7 (hauteur nasale) entraîne une augmentation significative de la distance; chez les hommes, il s'y ajoute encore les caractères 2 (longueur des jambes) et 8 (largeur nasale). Ces conclusions n'ont cependant qu'une valeur relative du moment qu'elles dépendent, jusqu'à un certain point, de l'ordre dans lequel on fait successivement entrer les différents caractères dans la discrimination (cf. § 2.43).

Pour des raisons qui tiennent au mode de construction des discriminations successives — c'est-à-dire utilisant un nombre croissant de caractères — il est avantageux de disposer

les caractères dans un ordre tel que, dans la suite des variations $D_p^2 - D_{p-1}^2$ les accroissements significatifs, au sens du test 240, apparaissent en premier, suivis par les accroissements tous fortuits. L'ordre de succession 7-4-1-2-8-5-3-6 pour les hommes, et 7-4-1-2-3-6-8-5 pour les femmes sont deux parmi ceux qui jouissent de cette propriété. Les résultats correspondants figurent au tableau 21. On constate qu'au cours de cette permutation, les caractères 6 (hauteur faciale) et 1 (buste) échangent leur rôle.

TABLEAU 21

Distances D_p^2 , accroissements $D_p^2 - D_{p-1}^2$ et test F relatif à ces accroissements (form. 240) pour deux discriminations raciales; ordre d'inclusion optimum

p	q	Discrimination ♂				Discrimination ♀			
		Caractères inclus	D_p^2	$D_p^2 - D_{p-1}^2$	F	Caractères inclus	D_p^2	$D_p^2 - D_{p-1}^2$	F
1	1	7	0.3283	...	...	7	0.3282	...	...
2	1	4	0.9221	0.5938	23.771 **	4	0.9961	0.6679	23.469 **
3	1	1	1.0603	0.1382	4.941 *	1	1.1548	0.1587	4.896 *
4	1	2	1.4149	0.3546	12.323 **	2	1.2453	0.0905	2.701 °
5	1	8	1.5349	0.1200	3.916 *	3	1.2786	0.0333	0.973 °
6	1	5	1.5738	0.0389	1.240 °	6	1.3063	0.0277	0.800 °
7	1	3	1.5829	0.0091	0.287 °	8	1.3066	0.0003	0.009 °
8	1	6	1.5866	0.0037	0.116 °	5	1.3071	0.0005	0.014 °

° valeur non significative.
 * valeur significative ($P = 0.05$).
 ** valeur hautement significative ($P = 0.01$).

On peut maintenant comparer les distances, basées sur les caractères correspondant aux accroissements $D_p^2 - D_{p-1}^2$ significatifs, à la distance D_8^2 selon la formule 237 dans laquelle $p = 5$ (caractères 7, 4, 1, 2 et 8), $q = 3$ (caractères 5, 3 et 6) pour les hommes, $p = 3$ (caractères 7, 4 et 1) et $q = 5$ (caractères 2, 3, 6, 8 et 5) pour les femmes. On trouve alors:

$$F(3; 202) = 0.544^\circ (\text{♂}); \quad F(5; 171) = 0.888^\circ (\text{♀}).$$

Cela signifie que les caractères 5, 3 et 6 chez les hommes, les caractères 2, 3, 6, 8 et 5 chez les femmes n'augmentent pas significativement la distance lorsqu'ils sont adjoints aux autres, qu'ils sont, en d'autres termes, de moindre importance pour la discrimination (cf. tabl. 22, § 2.42).

On peut enfin envisager la discrimination raciale de tous les individus sans égard au sexe, les caractères étant pris dans l'ordre 7-4-1-2 etc., avec:

$$N_1 = 144, N_2 = 277, N = 391, n = 389, \eta = 80.762 \ 148, \nu = 0.207 \ 615$$

(cf. tabl. 27, § 3.5). Les résultats figurent au tableau 22 (§ 2.42). Le test selon la formule 237 montre que la distance D_8^2 n'est pas significativement supérieure que D_4^2 . En effet:

$$F(4; 382) = 1.068^\circ.$$

Les caractères 7 (hauteur nasale), 4 (diamètre transverse), 1 (buste) et 2 (longueur des jambes) peuvent donc être considérés comme prédominants dans cette discrimination raciale commune.

Dans leur comparaison des méthodes sérologiques et somatométriques appliquées à la différenciation de certains groupes raciaux est-africains, EAST et OSCHINSKY ont appliqué un critère tout à fait semblable pour la désignation des huit variables importantes parmi les quatorze introduites à l'origine (EAST et OSCHINSKY, 1958, p. 41. Cf. aussi MUKHERJEE, 1951, p. 54).

2.42 Fonctions discriminantes simplifiées.

Par simplification, nous entendons ici la réduction du nombre de caractères utilisés pour la discrimination raciale. L'examen approfondi des distances associées à des discriminants dépendant d'un nombre variable de caractères a montré, en effet, quels étaient parmi ces derniers ceux qui introduits à la suite des autres, n'apportaient qu'une contribution négligeable à la distance et qui pourront dorénavant légitimement être omis. Par là, ce procédé illustre bien l'un des aspects principaux de la statistique, à savoir celui de la réduction des données à quelques caractéristiques peu nombreuses mais contenant l'essentiel de l'information que l'on peut extraire des données. Par ailleurs, la maniabilité d'un discriminant sera d'autant plus grande que le nombre des caractères qu'il utilise est petit, cela va sans dire. La réduction du nombre des caractères représente donc aussi un avantage pour l'application pratique.

On peut ainsi construire, pour la discrimination masculine, une fonction des cinq caractères 7, 4, 1, 2 et 8; pour la discrimination féminine une fonction des trois caractères 7, 4 et 1. Ce sont les caractères qui, pris dans cet ordre, augmentent significativement la distance D^2 . A titre de comparaison, et pour les besoins de la suite, nous avons construit pour les deux sexes les discriminations avec les trois caractères 7, 4 et 1, ainsi que celle avec les quatre caractères 7, 4, 1 et 2. Les coefficients h_k sont reproduits au tableau 22; ils ont été calculés selon le schéma du tableau 12 (§ 1.62), c'est-à-dire en complétant le calcul des distances D_p^2 du paragraphe précédent (§ 2.41) par la partie relative à la détermination des fonctions successives H_p (partie inférieure gauche des schémas du tableau 12).

La comparaison des deux discriminations raciales à quatre caractères au moyen du test approprié (form. 224 ou 228) permet d'en affirmer l'équivalence. En effet:

$$\begin{aligned} F(4; 384) &= 25.574 ** \quad (\text{signes } +) , \\ F(4; 384) &= 0.412 ^\circ \quad (\text{signes } -) . \end{aligned}$$

Il en est de même pour les deux discriminations raciales à trois caractères:

$$\begin{aligned} F(3; 385) &= 29.035 ** \quad (\text{signes } +) , \\ F(3; 385) &= 0.056 ^\circ \quad (\text{signes } -) . \end{aligned}$$

Pour ces quatre tests, on a pour constantes (cf. tabl. 27, § 3.5):

$$N_1 = 151, N_2 = 60, N_3 = 126, N_4 = 54, n = 387, \psi = 1/\sum_i (1/N_i) = 20.102\,842$$

et $p = 4$ ou 3 suivant le cas. La matrice $[w^{kl}]$ pour $p = 3$ a été calculée directement d'après la relation:

$$(328) \quad [w^{kl}] = ||w_{kl}||^{-1} \text{adj} [w_{kl}] ,$$

où $||w_{kl}||$ est le déterminant de la matrice $[w_{kl}]$, et $\text{adj} [w_{kl}]$ la matrice adjointe dont les éléments sont donc les cofacteurs du déterminant $||w_{kl}||$ (mineurs munis du signe approprié) (cf. AITKEN, 1949, pp. 39 et 53). La matrice inverse $[w^{kl}]$ pour $p = 4$ a été obtenue à partir de la précédente en introduisant une nouvelle variable (cf. COCHRAN, 1938, p. 173). Les quantités w_{kl} peuvent être reconstruites à partir des tableaux 27, 28 et 29 (§ 3.5), et les matrices $[(d'_k \pm d''_k) (d'_l \pm d''_l)]$ extraites du tableau 33 (§ 3.5). Les résultats de ces inversions figurent au tableau 35 (§ 3.5).

TABLEAU 22

Coefficients h_k , distances D_p^2 ($p = 3, 4, 5$ et 8) et test F relatif à l'accroissement de la distance (form. 237) pour différentes discriminations raciales réduites

	Discrimination						
	masculine			féminine		commune	
	$p = 3$	$p = 4$	$p = 5$	$p = 3$	$p = 4$	$p = 3$	$p = 4$
h_7	— 0.1648	— 0.1783	— 0.1852	— 0.2022	— 0.2117	— 0.1724	— 0.2011
h_4	0.1423	0.1453	0.1368	0.1615	0.1702	0.1555	0.1496
h_1	— 0.1124	— 0.2157	— 0.2158	— 0.1423	— 0.1507	— 0.1157	— 0.1827
h_2	...	0.1486	0.1413	...	0.0732	...	0.0965
h_8	...	...	0.1346	..	...	...	...
D_p^2	1.0603	1.4149	1.5349	1.1548	1.2453	1.0949	1.2887
D_8^2	1.5866	1.5866	1.5866	1.3071	1.3071	1.3570	1.3570
F	3.587 **	1.380 °	0.544 °	0.888 °	0.432 °	3.387 **	1.068 °
f_1	5	4	3	5	4	5	4
f_2	202	202	202	171	171	382	382

° valeur non significative.
 * valeur significative ($P = 0.05$).
 ** valeur hautement significative ($P = 0.01$).

L'équivalence des discriminations masculine et féminine tant pour $p = 3$ que pour $p = 4$ suggère aussitôt le calcul de la discriminante commune à trois ou quatre caractères (tabl. 22). En ce qui concerne le premier de ces discriminants communs, on peut enfin se demander si une approximation:

$$(329) \quad G_3 = -3x_1 + 4x_4 - 4x_7.$$

pourrait être utilisée en tant que simplification extrême du discriminant (commun) exact:

$$(330) \quad H_3 = -0.1157 x_1 + 0.1555 x_4 - 0.1724 x_7.$$

Comparons donc D_G^2 à D_H^2 en appliquant le test du paragraphe 1.42 (eq. 214 ou 215). D'après les données des tableaux 27, 28 et 29 (§ 3.5), ou bien plus précisément en partant des différences des moyennes de caractères entre WI et RI (sans égard au sexe), ainsi que des éléments w_{kl} du système des équations:

$$[w_{kl}][h_k] = [d_l]$$

pour les coefficients h_k de la fonction discriminante commune (§ 2.31), on aura:

$$D_G^2 = (\bar{G}_1 - \bar{G}_2)^2 / \text{var}(G) \\ = (27.9110)^2 / 683.8575 = 1.0907, \quad \text{tandis que (tabl. 22):}$$

$$D_H^2 = 1.0949, \quad \text{avec:}$$

$$N_1 = 277, N_2 = 114, N = 391, n = 389, p = 3 \quad (\text{tabl. 27, § 3.5}), \text{ de sorte}$$

que finalement:

$$F = (2; 387) = 0.138^\circ.$$

En définitive, la fonction G_3 (eq. 329 ci-dessus) des trois caractères 1 (buste), 4 (diamètre transverse) et 7 (hauteur du nez) donne une séparation pratiquement aussi bonne entre Walsers I et Romanches I (sans égard au sexe) que le meilleur discriminant racial H_3 (eq. 330 ci-dessus) basé sur les trois mêmes caractères et déterminé par les données observées.

2.43 Le rôle des corrélations simples entre caractères.

Le schéma de construction (tabl. 12, § 1.62) des discriminations utilisant un nombre croissant de caractères, lorsqu'on l'examine de près, montre que les coefficients $h_{p,k}$ ($k = 1, \dots, 8$ et $p = 1, \dots, 8$ successivement) et par conséquent les distances D_p^2 correspondantes dépendent de l'ordre de succession des caractères. A la figure 4 sont représentés, séparément pour les deux sexes, les accroissements $D_p^2 - D_{p-1}^2$ dans les discriminations raciales complètes, pour différents ordres de succession; $D_0^2 = 0$ par convention.

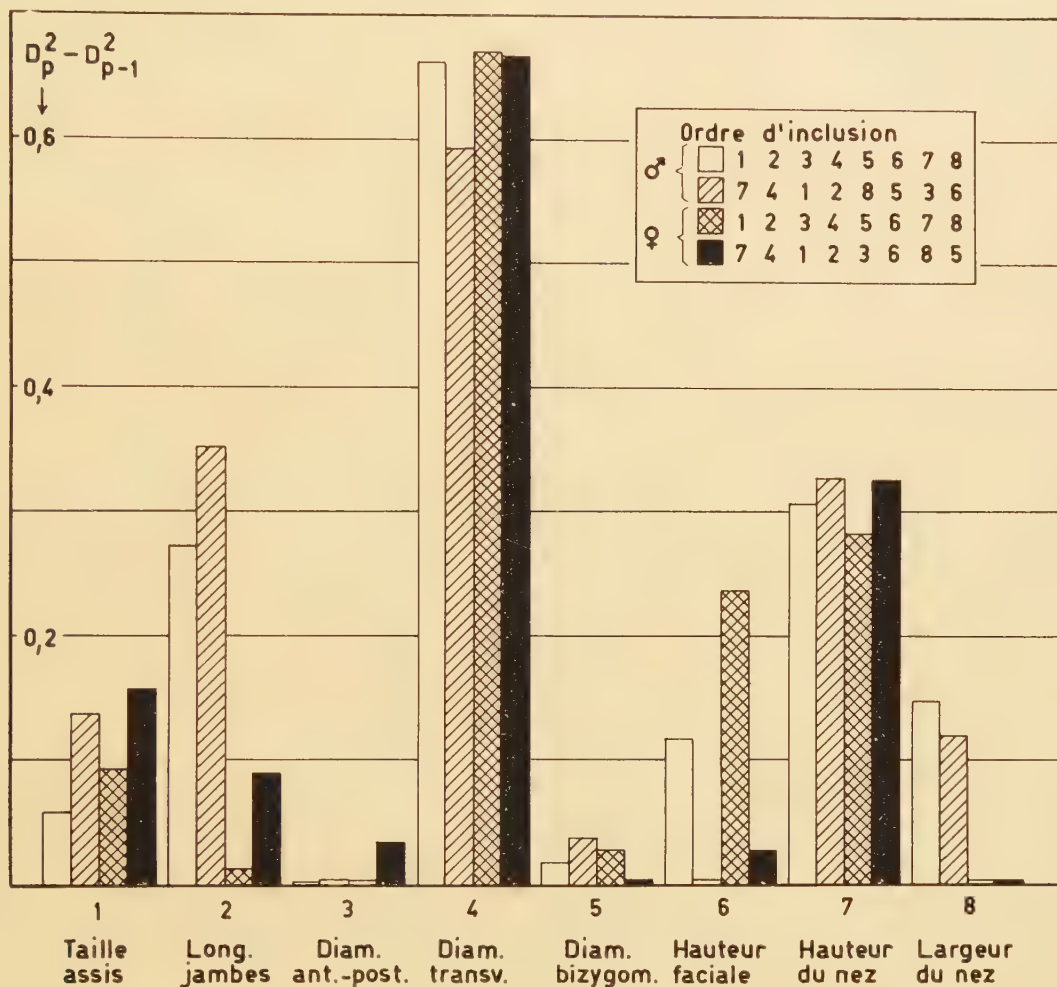


FIG. 4.

Répercussion, sur l'accroissement de la distance, d'une permutation dans l'ordre d'inclusion des divers caractères (discrimination raciale).

A part que ce diagramme reflète clairement l'importance relative des divers caractères, il fait ressortir les modifications des accroissements dues au changement d'ordre, et le cas du caractère 6 est remarquable à cet égard. Cela n'a, toutefois, rien de surprenant. Car en effet, dans l'arrangement primitif 1-2-3-4-5-6-7-8 (voir tabl. 20, § 2.41; ou bien § 2.11), le caractère 6 est placé avant le caractère 7 avec lequel il présente une corrélation simple hautement significative; après la corrélation des caractères 4 et 5, celle des caractères 6 et 7 est, chez les deux sexes, même la plus forte qui existe dans le matériel de cette enquête.

TABLEAU 23

Coefficients de détermination $B(x_k; x_l)$ des corrélations simples (à l'intérieur des sous-populations WI et RI); test F relatif à cette détermination (form. 332/333)

Corrélations		♂		♀	
Premier caractère (k) Second caractère (l)	kl	$B(x_k; x_l)$	Rang	$B(x_k; x_l)$	Rang
Taille assis					
Longueur des jambes	12	0.2500 **	3	0.0037 °	23
Diamètre antéro-postérieur . . .	13	0.0749 **	6	0.0814 **	6
Diamètre transverse	14	0.0309 *	13	0.0203 °	17
Diamètre bizygomatique	15	0.0563 **	8	0.0321 *	15
Hauteur faciale	16	0.0795 **	5	0.0477 **	12
Hauteur du nez	17	0.0005 °	28	0.0020 °	27
Largeur du nez	18	0.0057 °	22	0.0024 °	25
Longueur des jambes					
Diamètre antéro-postérieur . . .	23	0.0665 **	7	0.0022 °	26
Diamètre transverse	24	0.0044 °	24	0.0160 °	19
Diamètre bizygomatique	25	0.0340 **	12	0.0012 °	28
Hauteur faciale	26	0.1141 **	4	0.0520 **	11
Hauteur du nez	27	0.0068 °	20	0.0072 °	22
Largeur du nez	28	0.0125 °	17	0.0107 °	21
Diamètre antéro-postérieur					
Diamètre transverse	34	0.0394 **	10	0.0846 **	5
Diamètre bizygomatique	35	0.0230 *	15	0.0682 **	8
Hauteur faciale	36	0.0363 **	11	0.0899 **	3
Hauteur du nez	37	0.0062 °	21	0.0162 °	18
Largeur du nez	38	0.0099 °	18	0.0673 **	10
Diamètre transverse					
Diamètre bizygomatique	45	0.3536 **	1	0.4196 **	1
Hauteur faciale	46	0.0055 °	23	0.0362 *	14
Hauteur du nez	47	0.0032 °	25	0.0291 *	16
Largeur du nez	48	0.0232 *	14	0.0680 **	9
Diamètre bizygomatique					
Hauteur faciale	56	0.0139 °	16	0.0750 **	7
Hauteur du nez	57	0.0015 °	26	0.0364 *	13
Largeur du nez	58	0.0551 **	9	0.0870 **	4
Hauteur faciale					
Hauteur du nez	67	0.2716 **	2	0.2654 **	2
Largeur du nez	68	0.0006 °	27	0.0036 °	24
Hauteur du nez					
Largeur du nez	78	0.0078 °	19	0.0147 °	20
Limite de signification	$B_{0.05}$	0.0184 (♂)		0.0213 (♀)	
Limite de signification	$B_{0.01}$	0.0315 (♂)		0.0369 (♀)	

° détermination non significative.
 * détermination significative ($P = 0.05$).
 ** détermination hautement significative ($P = 0.01$).

Etant donné que le caractère 7 est très important pour la discrimination raciale, il fallait s'attendre à ce que, en son absence, un caractère étroitement lié à lui prît de l'importance, tel le caractère 6 dans D_6^2 pour l'ordre de succession primitif. Mais on devra voir cette importance diminuer si le caractère 7 précède 6 comme dans l'arrangement 7-4-1-2, etc.

L'apport d'un caractère nouvellement introduit à l'augmentation de la distance dépend donc, dans une certaine mesure, des corrélations de celui-ci avec les caractères déjà inclus, et cela conduit à la recherche d'un procédé établissant le meilleur ordre d'inclusion. Plus précisément, ce procédé devrait permettre de désigner d'avance, dans l'ensemble des caractères disponibles (relevés dans une enquête), ceux qui seront importants pour la discrimination; d'avance, cela veut dire: sans qu'il soit nécessaire d'effectuer la discrimination, si possible. Une telle méthode a été proposée par BARNARD, consistant à tester l'écart des moyennes, dans deux sous-populations comparées, du caractère nouvellement introduit lorsqu'on a éliminé, au moyen d'une analyse de régression linéaire multiple, l'influence des caractères précédemment inclus; la première variable choisie est celle qui conduit à la plus grande valeur de la statistique testant l'écart des moyennes des caractères dans deux sous-populations, en l'occurrence une valeur t (cf. BARNARD, 1934-35, p. 355). Cette méthode a l'inconvénient des procédés constructifs, à savoir qu'à mesure qu'on introduit de nouveaux caractères, les opérations deviennent de moins en moins praticables. En outre, on ne dispose en fin de compte ni des coefficients h ni de la distance D^2 qui permettrait de porter un jugement sur la qualité de la discrimination en cours.

Aussi nous semble-t-il plus avantageux — bien qu'à première vue tout aussi encombrant quant au calcul numérique — d'établir une fois pour toutes la discrimination utilisant tous les caractères disponibles, et de partir d'elle pour aboutir à des discriminations simplifiées par l'application des tests précédemment exposés. Il est incontestable que l'obstacle principal est constitué par l'inversion de la matrice $[w_{kl}]$ d'ordre p ($p = 8$ dans notre cas). Mais cette opération une fois effectuée, l'application des formules d'ajustement permet de calculer rapidement les distances D_{p-1}^2 — par exemple au moyen de la relation 273 — et, le cas échéant, les coefficients $h_{p-1,k}$ — par exemple par les relations 250 — pour des valeurs $p - 1$ décroissantes. Ces formules d'ajustement sont considérablement plus simples dans le processus d'exclusion d'une variable que ne le seraient les formules analogues pour l'inclusion d'un nouveau caractère.

Nous n'entrerons pas dans le détail de ces méthodes. Remarquons simplement que le comportement des divers caractères peut, jusqu'à un certain point, être prédit par l'examen attentif des corrélations (simples) entre caractères. Le tableau 23 indique, séparément pour les deux sexes, la valeur du coefficient de détermination $B(x_k; x_l)$ défini, pour chaque couple x_k, x_l de caractères, par les relations:

$$(331) \quad B(x_k; x_l) = (w_{kl})^2 / (w_{kk} \cdot w_{ll}), \quad \text{avec:}$$

$$nw_{kl} = \sum_{i=1}^2 \sum_{j=1}^{N_i} (x_{ijk} - \bar{x}_{ik})(x_{ijl} - \bar{x}_{il}), \quad \text{et:}$$

$$n = N_1 + N_2 - 2 \quad (i = 1: \text{WI}; i = 2: \text{RI}), \quad (\text{cf. } \S 1.12).$$

Le test de signification a été appliqué selon la relation:

$$(332) \quad B(N-2)/(1-B) = F(1; N-3),$$

fournissant la limite de signification, au seuil P , suivante:

$$(333) \quad B_P = F_P / (N - 3 + F_P) \quad (\text{cf. LINDER, 1957, p. 105}).$$

La figure 5 donne une représentation schématique de l'ensemble de ces corrélations. Le trait épais indique la détermination hautement significative ($P = 0.01$), le trait fin la

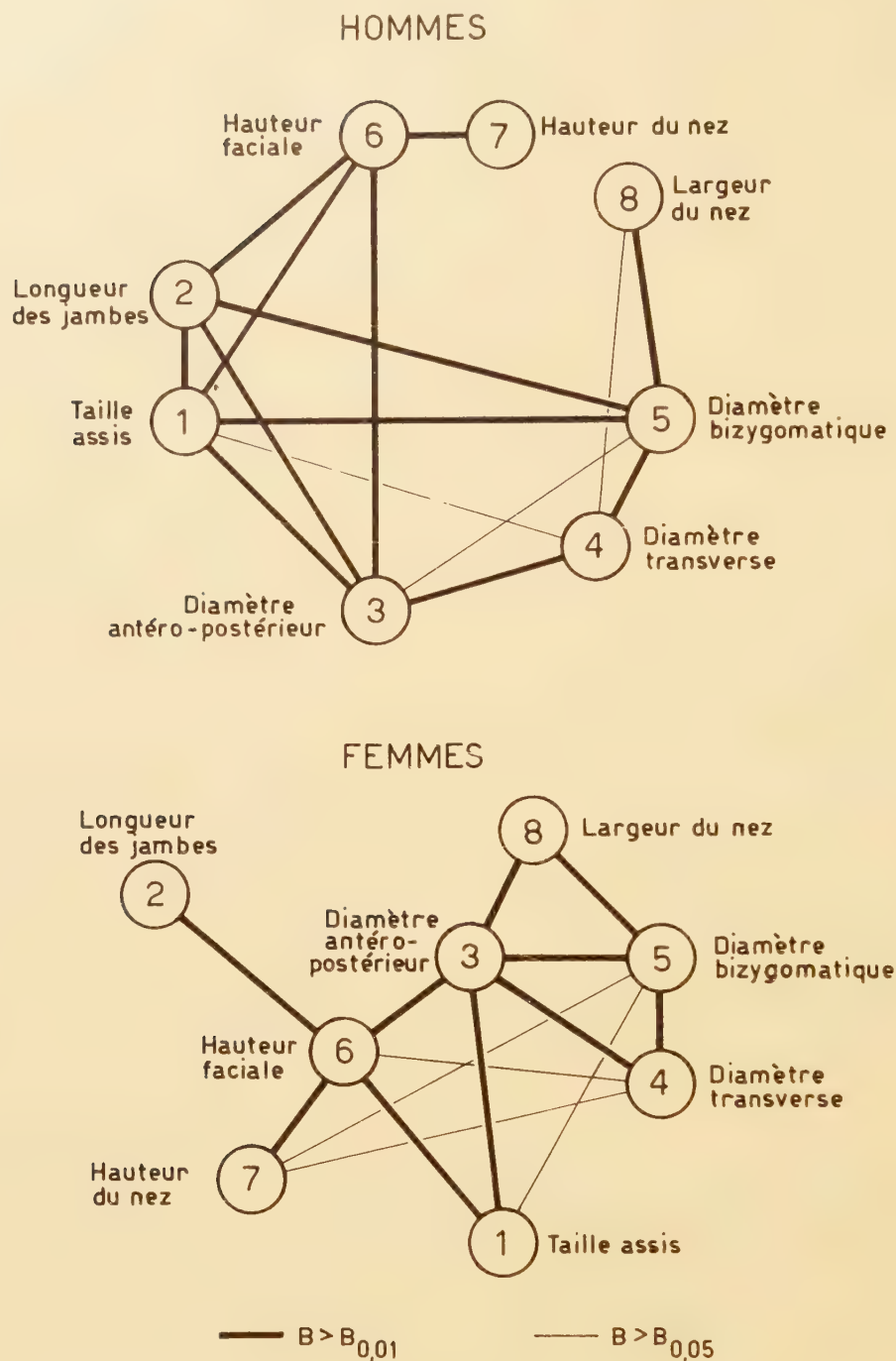


FIG. 5.

Corrélations simples entre caractères, à l'intérieur des sous-populations de type I.

détermination significative ($P = 0.05$), tandis que deux caractères non reliés ne présentent pas de corrélation significative entre eux; très approximativement, les caractères ont été placés d'autant plus proches les uns des autres que la corrélation entre eux est forte. Des caractères qui, tels 3, 5 ou 6, présentent de fortes corrélations avec un grand nombre d'autres caractères, n'auront vraisemblablement que peu d'importance dans la discrimination, tandis que 1, 2, 4 ou 7 sont au contraire relativement indépendants — notamment lorsqu'on fait entièrement abstraction de 3, 5 et 6 — et par conséquent prépondérants pour la discrimination. Le comportement du caractère 6, en particulier, s'explique aisément à l'aide de ces corrélations, et la supériorité de 7 par rapport à lui apparaît clairement.

De telles considérations peuvent se révéler d'une très grande utilité pratique, par exemple pour l'établissement provisoire d'un ordre de succession des caractères, ou bien pour le choix des caractères entrant dans une discrimination préliminaire n'utilisant qu'un nombre restreint de variables.

2.44 *Remarques sur une représentation géométrique.*

Aux figures 1 et 2 (§ 2.22) nous avons donné une représentation géométrique de la répartition des trois caractères buste (1), diamètre transverse (4) et hauteur du nez (7). Basée sur les moyennes des caractères dans les localités, elle était destinée à montrer jusqu'à quel point la séparation des points représentatifs en deux agglomérations, l'une walsers, l'autre romanche, se reflétait déjà dans les données observées, donc avant la discrimination à proprement parler.

Les considérations qui nous ont conduit à une telle représentation sont les suivantes: il est naturel d'associer aux p caractères mesurés sur un individu déterminé, les p coordonnées d'un point de l'espace à p dimensions. Aux deux sous-populations d'individus correspondraient, dans cet espace, deux nuages de points; on s'attendrait à une certaine correspondance entre la distance géométrique, au sens habituel, entre les centres de ces deux nuages, et la distance généralisée D_p^2 entre les deux sous-populations (cf. MAHALANOBIS et RAO, 1949, p. 116). En outre, les p coefficients h_k de la fonction discriminante correspondant à cette subdivision des données observées, pourraient être considérés comme définissant une direction, dans cet espace, le long de laquelle la séparation des deux sous-ensembles de points serait la meilleure possible (FISHER, 1938, p. 381). On aurait donc ainsi une interprétation géométrique immédiate tant pour la fonction discriminante que pour la distance qui lui est associée.

Il y a toutefois deux restrictions à observer à cette interprétation: l'analogie géométrique est correcte seulement si les p caractères sont indépendants (au sens du calcul des probabilités), et elle n'est commodément exploitable, pour une représentation spatiale parlante, que si $p = 3$ ou mieux encore $p = 2$. Ainsi s'explique pourquoi nous avons dû nous limiter, pour la représentation graphique, à trois caractères; et leur choix est justifié du fait qu'ils sont reliés par des corrélations (simples) qui ne sont pas très importantes (cf. tabl. 23, § 2.41).

A la figure 6 nous avons reporté les moyennes pour Walsers et Romanches des deux types et des deux sexes, avec les directions (en projection seulement) des deux discriminations raciales basées sur les trois caractères en question.

On pourrait être tenté de penser que la différence raciale — celle entre Walsers et Romanches — ne se manifeste pas de la même manière chez les sujets du type I que chez ceux du type II. Ce qui ressort clairement, en tout cas, c'est le parallélisme des discriminations masculine et féminine; l'équivalence de ces deux fonctions a, en effet, été établie (cf. § 2.42). La figure 6 donne, enfin, une image également très claire des différences sexuelles.

2.5 LES VALEURS INDIVIDUELLES

2.51 *Le cas particulier du type II.*

La fonction discriminante peut être utilisée pour déterminer l'appartenance, à l'une ou l'autre des deux sous-populations discriminées, de sujets dont on connaît toutes les mesures — en particulier celles qui apparaissent dans cette fonction — mais dont ces dernières n'ont pas servi à établir ladite discrimination. Nous avons ainsi la possibilité d'approfondir la question de la distinction typique par l'intermédiaire de la discrimination raciale. Comme nous avons considéré les sujets du type I comme plus représentatifs (cf. § 2.13), aussi bien chez les Walsers que chez les Romanches, il convient maintenant de préciser la position des Walsers II et des Romanches II par rapport à eux.

Utilisant pour simplifier le discriminant racial commun réduit aux quatre caractères les plus importants (cf. tabl. 22, § 2.42), on peut calculer la valeur individuelle de chaque sujet, sans égard au sexe, et représenter à la suite d'une classification adéquate les quatre sous-populations WI, WII, RI et RII sous forme de polygones de fréquences cumulées relatives (en échelle probit). C'est ce que représente la figure 7 permettant de conclure que le type II occupe une position intermédiaire entre les deux sous-populations de type I, à l'instar d'une population qui, anthropométriquement parlant, ne serait ni spécifiquement walser ni exclusivement romanche tout en étant proche des deux populations de référence. Examinons à l'aide d'un test approprié cette plus ou moins grande proximité (cf. § 1.32). Rappelons que nous négligeons la subdivision sexuelle dans toutes ces comparaisons.

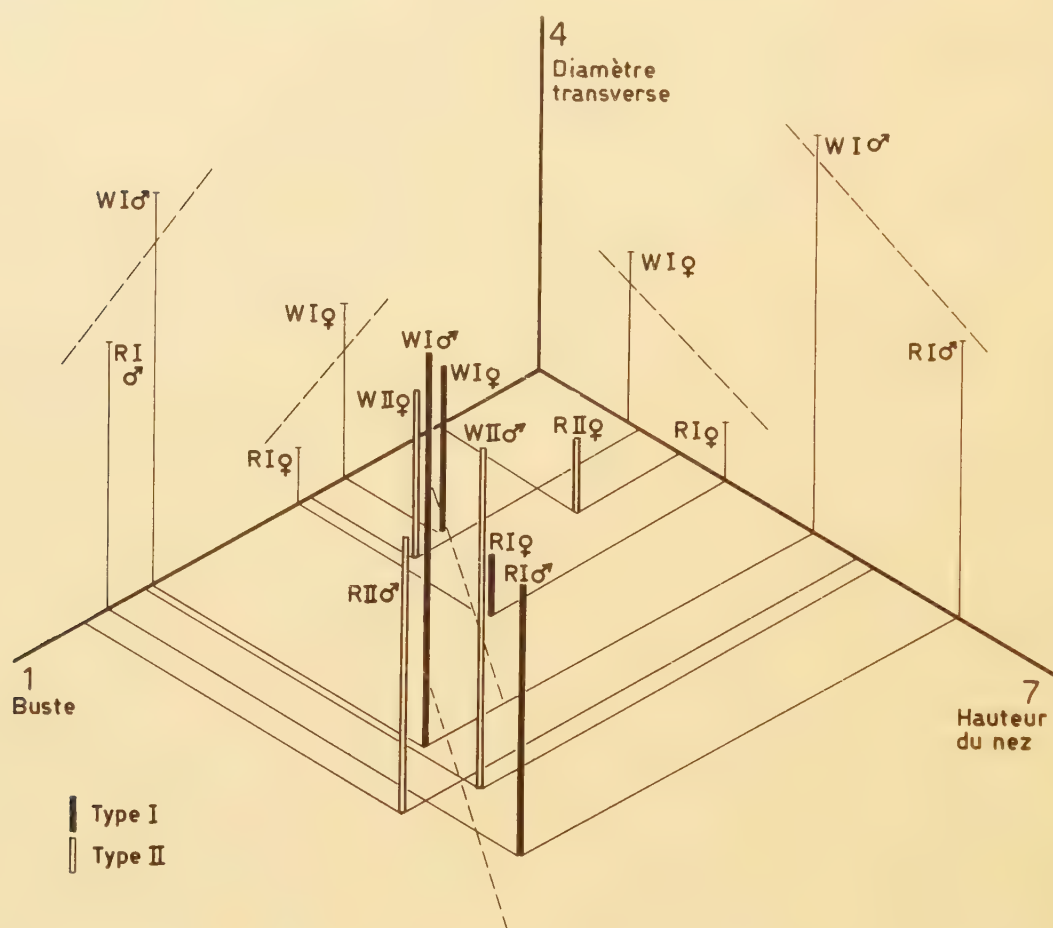


FIG. 6.

Points moyens des sous-populations principales (type I) en fonction des trois caractères buste, diamètre transverse, hauteur du nez, et directions définies par les discriminations raciales réduites.

Nous commencerons par les Walsers II ($i = 3$) que nous comparerons aux Walsers I ($i = 1$) et aux Romanches I ($i = 2$); le tableau 27 (§ 3.5) nous indique:

$$N_1 = 277, N_2 = 114, N_3 = 251, n = 639, \psi = 35.312\ 862, p = 4.$$

$$\sum_{k,l} w^{kl} (2\bar{x}_{3k} - \bar{x}_{1k} - \bar{x}_{2k}) (2\bar{x}_{3l} - \bar{x}_{1l} - \bar{x}_{2l}) = 0.449\ 860, \text{ de sorte que:}$$

$$F(4; 636) = 3.953^{**}.$$

La comparaison des Romanches II ($i = 3$) aux Walsers I ($i = 1$) et aux Romanches I ($i = 2$) donne:

$$N_1 = 277, N_2 = 114, N_3 = 38, n = 426, \psi = 8.500\ 135, p = 4;$$

$$\sum_{k,l} w^{kl} (2\bar{x}_{3k} - \bar{x}_{1k} - \bar{x}_{2k}) (2\bar{x}_{3l} - \bar{x}_{1l} - \bar{x}_{2l}) = 0.286\,190, \text{ de sorte que:}$$

$$F(4; 423) = 0.604^\circ.$$

Les éléments de l'inverse de la matrice:

$$[w_{kl}] = [\sum_{i,j} (x_{ijk} - \bar{x}_{ik})(x_{ijl} - \bar{x}_{il})]$$

ont été calculés comme ceux de la matrice $[w^{kl}]$ du paragraphe 2.42; les matrices $[w_{kl}]$ et $[w^{kl}]$ pour les deux comparaisons, figurent au tableau 36 (§ 3.5).

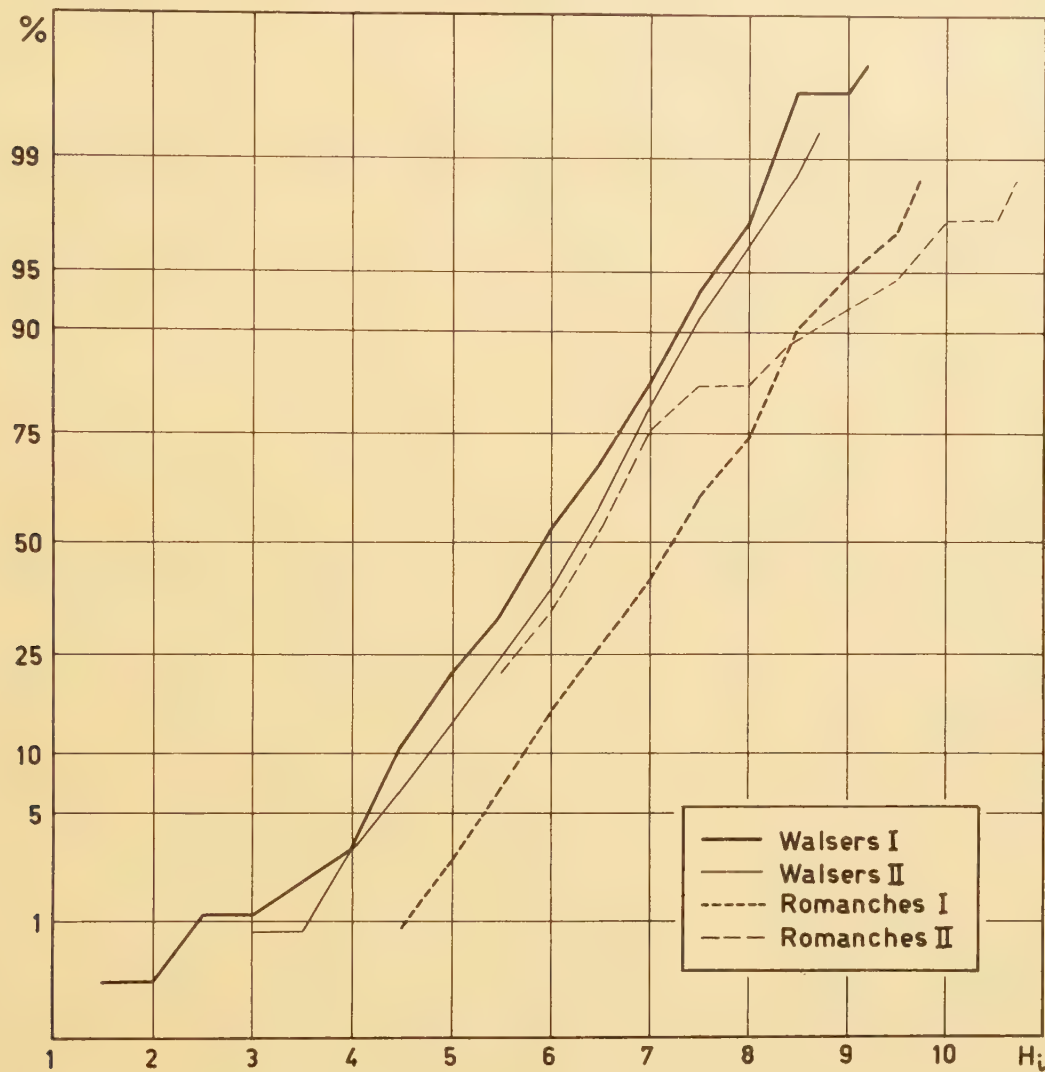


FIG. 7.

Fréquences cumulées des valeurs individuelles H_i dans les principales sous-populations de type I (échelle probit; discrimination raciale réduite commune).

Ces résultats montrent, comme le laissait supposer la figure 7, que les Walsers II ne sont pas également proches des Walsers I et des Romanches I. S'il subsistait un doute quant à la distance plus grande entre Walsers II et Romanches I, qu'entre Walsers II et Walsers I, il suffirait d'effectuer le calcul de ces deux distances pour les comparer ensuite. En ce qui concerne les Romanches, le test n'indique aucune inégalité des distances entre Romanches II et Romanches I d'une part, et celle entre Romanches II et Walsers I de l'autre.

La courbe représentative des Romanches présente, dans sa partie supérieure, un changement de direction assez prononcé chez le type II (voir fig. 7). Cette perturbation est en partie provoquée par les valeurs exceptionnellement grandes de la fonction discriminante

pour deux sujets de la localité romanche de *Tinizong*. En effet, l'exclusion de ces deux valeurs conduirait à une courbe représentative, pour les Romanches II, moins irrégulière et située en tout cas entièrement entre celle des Romanches I et des Walsers I; cette courbe n'est pas reproduite dans la figure 7. Notons encore que de telles perturbations sont encore exagérées par le fait des petits effectifs.

Les quantités $e_k = 2\bar{x}_{3k} - \bar{x}_{1k} - \bar{x}_{2k}$ (moyennes relatives à l'ensemble des deux sexes), calculées pour les deux comparaisons d'après les données du tableau 27 (§ 3.5), sont indiquées au tableau 37 (§ 3.5) ainsi que les produits $e_k e_l$.

2.52 Classification individuelle.

La valeur du discriminant racial commun réduit aux quatre caractères principaux ayant été calculée séparément pour chaque individu ainsi que pour les moyennes des localités, il fallait s'attendre à voir certaines de ces dernières occuper une position sinon assez écartée, du moins manifestement intermédiaire par rapport à la moyenne générale des Walsers I et des Romanches I (sans égard au sexe). Tel est le cas pour les localités de *Tinizong* (Ti), *Langwies* (Lw) par exemple, de même pour *Glas* (Gl) (cf. fig. 8).

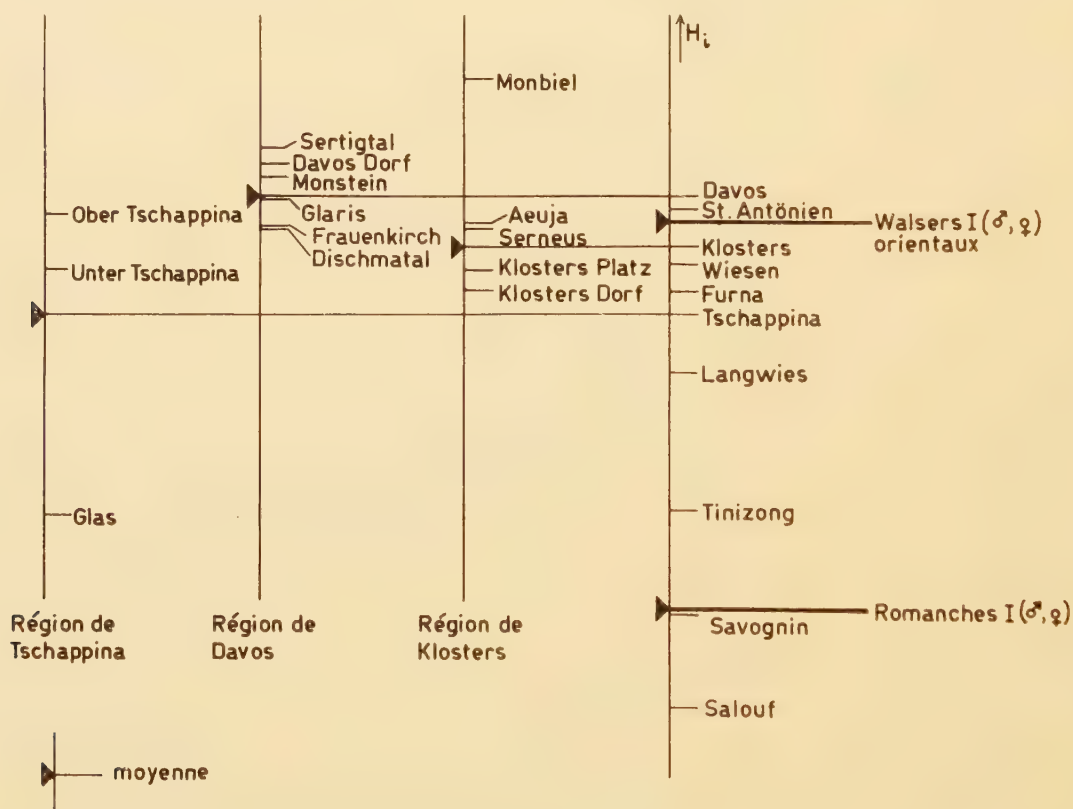


FIG. 8.

Discrimination des localités ou groupes de localités, selon la fonction discriminante réduite commune: moyennes des valeurs individuelles H_i (type I).

Nous allons examiner de près les quatre sujets mesurés à *Glas*, localité qui fait partie de l'agglomération de *Tschappina* considérée comme walser occidentale et restée écartée de ce fait des discriminations fondamentales du présent travail (cf. § 2.11). Le tableau 24 indique, entre autres, les mesures x_{3k} de ces quatre sujets ainsi que la valeur du discriminant montrant, plus nettement encore que la moyenne, le décalage dans la direction de la moyenne romanche. Appliquons un test (form. 150-152) pour savoir si l'attribution de ces quatre sujets à l'un des lots, Walsers I ou bien Romanches I, a un sens. La matrice de dispersion

(eq. 147) exigée pour ce test étant à prendre à l'intérieur des deux échantillons de comparaison — ici Walsers I d'une part, Romanches I d'autre part, sans égard au sexe — ses éléments seront simplement les grandeurs w_{kl} du système d'équations déterminant les h_k du discriminant racial commun basé sur les caractères 1, 2, 4 et 7 (cf. tabl. 34 ou 38, § 3.5). Son inverse, calculé comme il l'est indiqué au paragraphe 2.42, et les quantités utiles d'_k, d''_k (cf. tabl. 14, § 2.12) déterminent les valeurs F figurant au tableau 24. Aucune de ces quatre valeurs F n'est significative, ce qui indique que dans l'ensemble, les différences :

$$\left. \begin{aligned} d'_k &= x_{3k} - \bar{x}_{1k} \\ d''_k &= x_{3k} - \bar{x}_{2k} \end{aligned} \right\} \quad \text{pour } k = 1, 2, 4 \text{ ou } 7,$$

où $\bar{x}_{1k}, \bar{x}_{2k}$ désignent respectivement la moyenne générale (sans égard au sexe) des Walsers I et celle des Romanches I, ne doivent pas être considérées comme significativement différentes. En d'autres termes, les quatre sujets mesurés à *Glas* ne s'écartent pas davantage de la moyenne romanche I qu'ils ne le font de la moyenne walser I, si l'on considère les quatre caractères globalement. On ne peut donc pas classer ces quatre sujets sans ambiguïté dans l'un des ensembles Walsers I ou Romanches I.

TABLEAU 24

*Exemple de classification individuelle de quatre sujets walsers occidentaux
(test form. 150/152)*

Sujet (code)	Mesures				Discri- mination L	$e_k = \bar{x}_{3k} - \bar{x}_{1k} - \bar{x}_{2k}$				F^1
	x_{31}	x_{32}	x_{34}	x_{37}		d_1	d_2	d_4	d_7	
22101	41	33	52	53	7.185	7.299	8.963	0.299	1.705	0.323 °
22102	42	42	54	59	7.407	9.299	26.963	4.299	13.705	1.688 °
22301	39	32	57	54	6.370	3.299	6.963	10.299	3.705	0.180 °
22302	34	25	40	48	7.468	6.701	7.037	23.701	8.295	0.801 °
WI ²	36.964	29.090	53.982	51.155	6.158	...	...	...	...	...
RI ²	37.737	27.947	49.719	53.140	7.446	...	...	...	...	...

¹ limite de signification: $F(4; 385)_{0.05} = 2.396$.
² moyenne du type I (hommes et femmes).

Cette conclusion pourrait paraître surprenante. On s'attendrait, en effet, à ce que des sujets ayant des valeurs H_{ij} aussi voisines de la moyenne générale romanche que trois parmi ceux que nous avons examinées à *Glas*, soient sans autres attribuables à la sous-population romanche (cf. fig. 7, § 2.51). Il faut noter cependant que les deux sous-populations discriminées empiètent très fortement l'une sur l'autre en dépit de ce que la distance entre elles s'écarte significativement de zéro (voir fig. 9). Cela implique que l'intervalle des valeurs individuelles H_{ij} dont la classification est douteuse, est relativement grand et s'étend vraisemblablement bien au-delà des deux moyennes générales, celle des Walsers I et celle des Romanches I.

2.53 Tests directs utilisant les valeurs individuelles.

Une fonction discriminante ayant été mise au point, on pourrait en calculer la valeur individuelle H_{ij} pour chaque sujet mesuré et considérer l'ensemble de ces valeurs H_{ij} comme

nouvelles données numériques auxquelles on appliquerait les tests usuels de comparaison de moyennes. Cette manière de faire est justifiée, dira-t-on, du fait qu'elle ne fait pas intervenir le concept de distance avec ce que celle-ci implique de difficultés quant aux opérations matricielles, ne serait-ce que l'inversion. D'autre part, ajoutera-t-on, du moment que de par le principe même de la méthode, on remplace les diverses mesures par une unique combinaison (linéaire), il est naturel de baser toute la suite de l'analyse précisément sur les valeurs de cette combinaison.

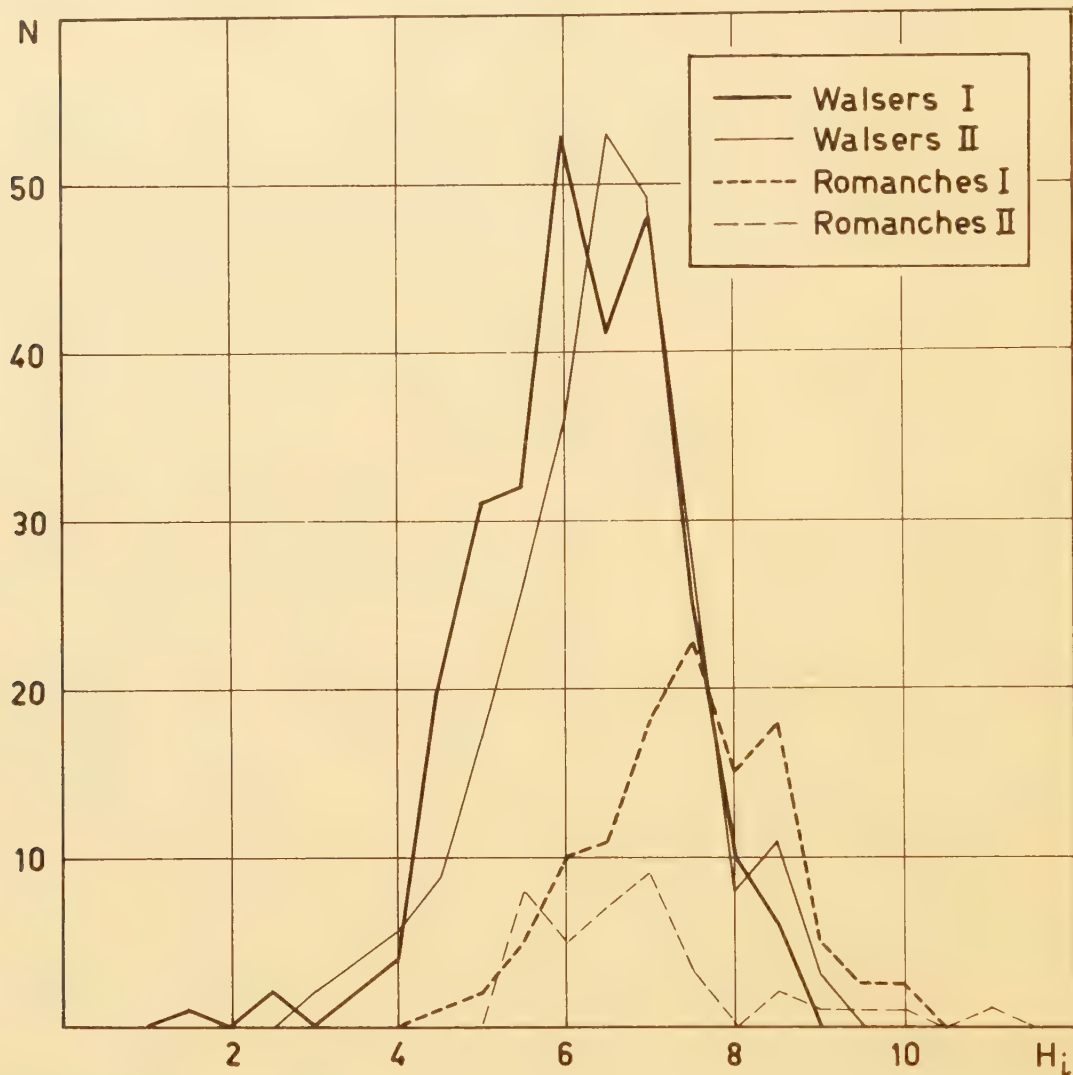


FIG. 9.

Répartitions de fréquence des valeurs individuelles H_i dans les principales sous-populations de type I (fonction discriminante réduite commune).

Considérons la discrimination raciale commune au moyen des quatre caractères principaux (tabl. 22, § 2.4), et ses valeurs individuelles que nous appellerons H_j , laissant tomber l'indice i , provisoirement. L'application, à ces valeurs H_j d'un test concernant l'écart de deux moyennes, par exemple, est légitime pour autant que soient vérifiées les hypothèses assurant la validité de ce test. S'il s'agit de tester la différence $\bar{H}' - \bar{H}''$ au moyen du test t selon la relation 51:

$$(334) \quad \frac{|\bar{H}' - \bar{H}''|}{s_H} \sqrt{\frac{N' \cdot N''}{N' + N''}} = t_{[f]} ,$$

avec $f = N' + N'' - 2$ degrés de liberté (N' , N'' étant les effectifs des deux échantillons à comparer), ou bien, ce qui est plus expéditif pour le calcul pratique, selon :

$$(335) \quad (\bar{H}' - \bar{H}'')^2 N' N'' / s_H^2 (N' + N'') = F(1; N' + N'' - 2) ,$$

ces hypothèses sont principalement les suivantes : les sous-ensembles de valeurs H_j doivent suivre des répartitions normales ayant même paramètre de dispersion. Dans la figure 7 (§ 2.51) les courbes représentatives des Walsers I ou II ainsi que celle des Romanches I montrent un tracé à peu près rectiligne et parallèle, indiquant par là qu'en première approximation ces trois sous-ensembles sont normaux avec même dispersion. Nous avons déjà examiné le comportement aberrant de la courbe des Romanches II (cf. § 2.51). En estimant d'après cela les hypothèses invoquées suffisamment vérifiées, on pourrait effectuer les tests des écarts de moyennes tels qu'ils sont reproduits schématiquement dans la figure 10 indiquant, pour chaque couple de moyennes — pour autant que cette comparaison soit sensée — la plus petite différence significative au seuil de 1 %, déterminée d'après les relations suivantes :

$$(336) \quad (\bar{H}' - \bar{H}'')_{0,01} = \sqrt{s_H^2 F(f_1; f_2) \cdot (N' + N'') / N' N''}$$

$$= s_H t_{[f_2]} \sqrt{(N' + N'') / N' N''} , \quad \text{avec :}$$

$$f_1 = 1; f_2 = N_{\text{WI}} + N_{\text{WII}} + N_{\text{RI}} + N_{\text{RII}} - 4 = N - 4 ;$$

s_H^2 étant ici le carré moyen à l'intérieur des quatre sous-ensembles WI, WII, RI et RII' déduit d'une analyse de variance simple des valeurs individuelles H_j (sans égard au sexe)' analyse non reproduite ici. Nous laissons au lecteur anthropologiste le soin d'exploiter les résultats de la figure 10. Remarquons que nous sommes ici en présence de quatre échantillons, de sorte que l'analyse de variance simple selon le tableau 5 (§ 1.31) et le test F correspondant (eq. 136) ne peuvent pas être appliqués. Nous n'avons en effet pas démontré que la généralisation de ce test, au cas de plus de deux échantillons, reste exacte.

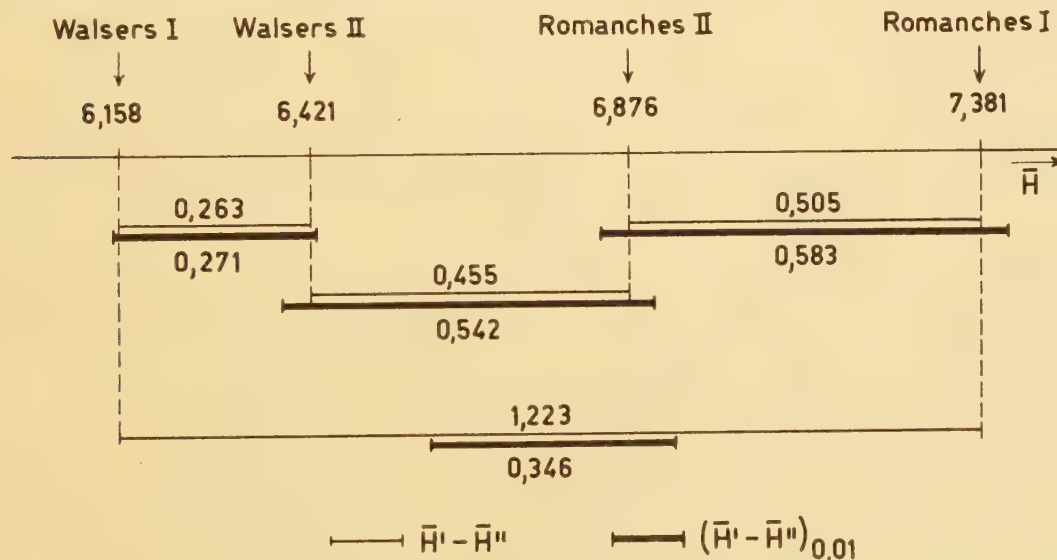


FIG. 10.

Centre des répartitions de valeurs individuelles H_i : écarts et plus petite différence significative entre les moyennes dans les principales sous-populations de type I ou II (fonction discriminante réduite commune).

Insistons toutefois sur le fait que l'on pourrait ne pas se contenter d'une simple vérification visuelle (graphique) des hypothèses assurant la validité du test en question. Il y aurait lieu, dans ce cas, d'établir d'une part la normalité des quatre sous-ensembles de valeurs H_j , l'égalité des dispersions d'autre part, par un test objectif.

La normalité, plus exactement le défaut de normalité, serait par exemple testé au moyen d'une mesure d'obliquité (dissymétrie) J_1 et une mesure de *kurtosis* J_2 , mesures basées toutes deux sur les estimations K des premiers cumulants de la répartition considérée (cf. FISHER, 1958, §§ 14 et 19):

$$(337) \quad J_1 = K_3/K_2 \sqrt{K_2}, \quad \text{et:}$$

$$(338) \quad J_2 = K_4/K_2^2.$$

On associerait à ces statistiques leur déviation-standard s_J qui sont données par les relations:

$$(339) \quad s_{J_1}^2 = 6N(N-1)/(N-2)(N+1)(N+3), \quad \text{et:}$$

$$(340) \quad s_{J_2}^2 = 24N(N-1)^2/(N-3)(N-2)(N+3)(N+5).$$

TABLEAU 25

Test de la normalité des répartitions de valeurs H_j dans les sous-populations WI, WII, RI, RII (form. 337/340); discrimination raciale réduite commune

		Sous-populations (♂, ♀)			
		WI	WII	RI	RII
Effectif	N	277	251	114	38
Cumulants estimés .	K_1	0.321	— 0.112	— 0.114	— 2.237
	K_2	5.272	4.712	5.071	12.135
	K_3	0.434	— 4.564	— 2.711	— 36.146
	K_4	17.210	8.603	— 2.883	— 205.187
Obliquité	J_1	0.036	— 0.446	— 0.237	— 0.855
Kurtosis	J_2	0.619	0.388	— 0.112	— 1.393
Ecart-types	s_{J_1}	0.146	0.154	0.226	0.383
	s_{J_2}	0.292	0.306	0.449	0.750
Tests $t_{ f }$ ($f = \infty$)	t_{J_1}	0.245 °	2.903 **	1.048 °	2.312 *
	t_{J_2}	2.123 *	1.266 °	0.249 °	1.859 °
° valeur non significative. * valeur significative ($P = 0.05$). ** valeur hautement significative ($P = 0.01$).					

Le quotient J/s_J serait une variable $t_{|f|}$ (degrés de liberté $f = \infty$), c'est-à-dire une variable normale standardisée u , lorsque l'effectif N est grand. Nous ne nous attarderons pas sur le détail du calcul des grandeurs K (cf. SNEDECOR, 1956, p. 202). Le tableau 25 indique le résultat de ces tests appliqués à notre matériel numérique, les valeurs H_j à savoir. La répartition WI est symétrique mais légèrement leptocurtique (valeurs centrales relativement trop fréquentes). La répartition WII est dissymétrique (sa médiane est supérieure à la moyenne arithmétique). La répartition RI ne présente ni dissymétrie ni *kurtosis*; enfin, la répartition RII est nettement dissymétrique comme nous le savons déjà, mais il n'y a pas lieu d'attacher une grande importance à cette dernière conclusion, l'effectif ($N = 38$)

étant petit (cf. aussi fig. 9, § 2.52). D'une façon générale, il n'y aurait donc que la répartition RI qui pourrait, au sens strict, être considérée comme normale relativement à ces deux critères.

En ce qui concerne par ailleurs l'égalité des quatre dispersions $s_{H,i}$ ($i = 1, 2, 3$ ou 4), le test de BARTLETT peut être appliqué, selon lequel :

$$(341) \quad nL_n\left\{\sum_i n_i s_{H,i}^2 / n\right\} - \sum_i \{n_i L_n s_{H,i}^2\} = \chi_{[f]}^2, \text{ avec:}$$

$$(342) \quad s_{H,i}^2 = \sum_{j=1}^{N_i} \{(H_{ij} - \bar{H}_i)^2 / n_i\}, \quad i = 1, \dots, m (m = 4); j = 1, \dots, N_i;$$

$$n_i = N_i - 1; n = \sum_i n_i; \text{ degrés de liberté } f = m - 1 (= 3).$$

L_n désigne le logarithme naturel. (Cf. LINDER, 1953, p. 28; BARTLETT, 1937.) Les opérations sont résumées au tableau 26. La valeur χ^2 est inférieure à la limite de signification, indiquant par là que les différences entre ces quatre dispersions sont attribuables au hasard. A cet égard, la comparaison des moyennes paraît irréprochable, alors que le défaut de normalité précédemment mis en évidence dans certaines des sous-populations envisagées, pourrait donner lieu à des objections principales. Sans vouloir entrer dans le détail de ces dernières, il convient néanmoins de remarquer qu'en raison des corrélations, les caractères entrant dans la fonction discriminante ne sont pas indépendants, qu'on ne peut par conséquent pas affirmer *a priori* que ce discriminant produira des valeurs réparties normalement, en dépit de ce que les composantes sont assez généralement considérées comme variables normales (cf. pourtant ELDERTON and WOO, 1932).

TABLEAU 26

*Test d'homogénéité des variances s^2 dans les sous-populations WI, WII, RI, RII
(form. 341/342); discrimination raciale réduite commune*

Sous-populations	i	N_i	n_i	$s_{H,i}^2$	$n_i s_{H,i}^2$	$L_n s_{H,i}^2$	$n_i L_n s_{H,i}^2$
WI (♂, ♀)	1	277	276	1.412	389.712	0.345	95.220
WII (♂, ♀)	2	252	251	1.314	329.814	0.273	68.523
RI (♂, ♀)	3	115	114	1.775	202.350	0.574	65.436
RII (♂, ♀)	4	38	37	1.694	62.678	0.527	19.499
Total	...	...	678	...	984.554	...	248.678
$\chi_{[f]}^2 = \{678 L_n (984.554/678)\} - 248.678 = 4.223 \quad (f = 3)$							
Limite de signification: $\chi_{0.05}^2 = 7.815 \quad (f = 3)$							

Rappelons, pour terminer, une hypothèse importante que nous avons admise et utilisée dès le départ: celle de l'équidispersion des caractères dans les diverses sous-populations discriminées (§ 1.11). En admettant cette hypothèse vérifiée, on devrait pouvoir en déduire sans autres, semble-t-il, l'égalité des dispersions des valeurs H_{ij} dans les sous-populations en question. Toute investigation visant à établir ou à tester cette égalité deviendrait ainsi superflue. Pour répondre à cette objection — dont nous ne contestons d'ailleurs pas du tout le bien-fondé — nous ferons remarquer que l'hypothèse de l'équidispersion n'a, dans le

présent travail, fait l'objet d'aucune analyse spéciale; sans doute que cela pourrait se faire, sans trop de complications, à l'aide des tables de PILLAY par exemple (PILLAY, 1957, table 16; cf. également ANDERSON, 1958, chap. 10). Cette hypothèse peut donc être mise en question, en principe, aussi bien que la provenance de nos échantillons de répartitions normales multidimensionnelles. Nous pensons qu'il y a là, pour le moins, matière à discussion.

3. RÉCAPITULATION

3.11 *Résumé.*

Partie théorique. — La partie théorique débute par une description algébrique des deux méthodes usuelles pour la détermination de la fonction discriminante linéaire (au sens de R. A. FISHER): condition d'extremum aboutissant à une équation déterminantale, analogie avec la régression linéaire multiple. Une relation entre ces deux méthodes est redémontrée et les analyses de variance relatives sont définies. Sont également introduits les concepts de distance généralisée selon MAHALANOBIS et le rapport généralisé de HOTELLING.

Les tests concernant la distance entre deux ou trois échantillons sont décrits formellement avec, comme cas particulier, l'application au problème du classement d'une observation unique dans l'un ou l'autre de deux échantillons.

La partie théorique se poursuit par la discussion des tests concernant la précision et la comparaison de fonctions discriminantes: comparaison d'une fonction calculée et d'une fonction désignée d'avance, comparaison de deux fonctions calculées.

Ces comparaisons conduisent aux considérations visant à la simplification de fonctions discriminantes: problèmes d'exclusion ou d'inclusion et importance relative des diverses variables participant à la discrimination, détermination de fonctions simplifiées ne comportant qu'un nombre restreint de variables; répercussion de ces simplifications sur la valeur de la distance.

La partie théorique se termine par une brève revue de la méthodologie arithmétique: résolution des systèmes d'équations linéaires et inversion matricielle par la méthode de condensation selon GAUSS-DOOLITTLE; indication des schémas de calcul abrégé avec contrôle incorporé.

Partie d'application. — Les tests précédemment redémontrés sont appliqués à l'élaboration systématique de relevés anthropométriques concernant certaines comparaisons de populations alpines aux *Grisons (Suisse)*. Deux groupes de sujets généalogiquement, linguistiquement et ethnologiquement discernables sont discriminés au moyen de huit caractères somatiques métriques: le buste (ou taille assis), la longueur des jambes (différence entre la stature et la taille assis), les diamètres antéro-postérieur et transverse maximums du crâne, le diamètre bizygomatique du crâne, la hauteur faciale (morphologique), la hauteur et la largeur du nez.

De l'ensemble des tests de comparaison, il résulte que tant chez les sujets masculins que féminins, la particularité anthropométrique de ces deux populations se manifeste de façon assez semblable et se traduit chez les deux sexes par l'existence d'une discrimination en tout cas significative.

Il résulte par ailleurs des tests de simplification que les caractères importants, dans ces discriminations, sont essentiellement le buste, la longueur des jambes, le diamètre transverse maximum du crâne et la hauteur du nez.

Ces deux résultats autorisent conjointement la construction d'une unique fonction de ces quatre caractères seulement, valable indifféremment pour les deux sexes cependant, fonction pouvant à son tour être approximée sans perte appréciable d'information par une forme linéaire simplifiée à coefficients tous entiers.

L'importance prépondérante de ces quatre caractères pour la discrimination en question, importance explicable par le rôle des dépendances intercaractérielles, se trouve en

effet être pleinement confirmée par l'analyse des valeurs de ces corrélations simples dans les couples de caractères.

Enfin, l'existence d'une fonction discriminante simplifiée indépendante du sexe a conduit à l'étude de la répartition des valeurs prises par cette fonction pour les individus des diverses sous-populations considérées en vue de situer ces dernières les unes par rapport aux autres, également en vue d'effectuer la classification individuelle d'une localité étrangère au domaine de discrimination.

3.12 *Summary.*

Theoretical part. — To start with, this part presents the two usual procedures for the computation of R. A. FISHER's linear discriminant function: the maximizing of the between to within variance ratio, resulting in a determinantal equation, and the analogy with multiple regression technique. The link between these two procedures is indicated as well as respective analyses of variance. The definition of MAHALANOBIS' generalized distance and HOTELLING's generalized ratio are introduced without proof.

Significance tests of the distances between two or three samples are then reconstructed and applied to the particular problem of classifying a single observation into one of two samples.

The theoretical review proceeds with the discussion of questions referring to the comparison or precision of discriminant functions: how to test the difference between an assigned and a calculated, or between two calculated discriminant functions.

The resulting comparison criterion leads to the distinction between useful variates to be included in, and insignificant one's to be excluded from the discriminatory analysis, thus solving the difficulty of reducing the number of variates which significantly improve the goodness of discrimination as measured by general distance.

At last, some self-checking computational schemes are indicated for the numerical solution of linear equations and matrix inversion by means of the GAUSS-DOOLITTLE condensation method.

Applications. — The above-mentioned tests are systematically applied to anthropometric data collected in the *Alpine Grisons (Switzerland)* in view of group comparison. Two genealogically, linguistically and ethnically distinct classes of individuals are discriminated by means of eight physical measurements: sitting height, leg length (difference between stature and sitting height), maximum head breadth and length, bizygomatic diameter, total morphological face height, nasal height and breadth.

From the comparison tests on the whole, it may be concluded that male and female individuals are quite alike regarding the anthropometric particularity of the two investigated classes as indicated by significant discrimination.

On the other hand, the relevancy of each character has been tested and allows the statement that for the discrimination under consideration, sitting height, leg length, head breadth and nasal length are the four most effective distance improving characters.

These conclusions form jointly the basis to the construction of a unique discriminant function depending only on the four important characters mentioned before and applying to either sex indifferently. The so reduced formula in turn is still capable of approximation, without notable loss of information, by a simpler linear compound whose coefficients are all whole numbers.

The prevalent importance of those four characters in the present discriminatory analysis may be expected to be due mainly to their interdependences. In fact, clear evidence of this assertion is shown by the inspection of the set of all possible simple correlations between the eight measured characters.

A simplified function having been computed independently on sexual differentiation, the distributions of the values for individuals belonging to several subpopulations have been worked out and applied to different group or individual classifications.

3.13 Zusammenfassung.

Theorie. — Der theoretische Abschnitt beginnt mit der algebraischen Einführung der beiden üblichen Bestimmungsmethoden für die FISHER'sche lineare Trennfunktion: Lösung der durch das Extremalverfahren bestimmten Determinantengleichung, Analogie der mehrfachen linearen Regression. Die Beziehung zwischen den beiden Methoden und diesbezügliche Streuungszerlegungen werden erläutert und die Begriffe des verallgemeinerten Abstandes von MAHALANOBIS sowie des HOTELLING'schen verallgemeinerten Verhältnisses ohne Beweis ebenfalls angeführt.

Prüfverfahren für den Vergleich von zwei, bzw. drei verallgemeinerten Abständen werden formal abgeleitet und am Problem des Einordnens einer Einzelbeobachtung in die eine von zwei Stichproben angewandt.

In der Folge befasst sich der theoretische Abschnitt mit Genauigkeits- und Ähnlichkeitstests im Gebiete des Trennverfahrens, wie zum Beispiel für die Vergleiche zweier errechneter Trennfunktionen, oder für den Unterschied zwischen einer vorgegebenen und einer errechneten Trennfunktion.

Diese Vergleiche liegen den darauffolgenden Ausführungen über die Vereinfachung von Trennverfahren zugrunde: Ein- und Ausschluss von Veränderlichen, relative Wichtigkeit und Auswahl der in das Trennverfahren einzubeziehenden Veränderlichen im Hinblick auf vorteilhafteste Verminderung der Variablenanzahl; Auswirkung dieser Vereinfachung auf den verallgemeinerten Abstand.

Der theoretische Abschnitt schliesst mit einer kurzen Uebersicht der hauptsächlichsten in der vorliegenden Arbeit zur Anwendung gelangten Rechenverfahren: numerisches Lösen von linearen Gleichungsscharen und Matrixkehrung mittels der Kondensationsmethode nach GAUSS-DOOLITTLE.

Anwendung. — Die soeben erläuterten Prüfverfahren bilden systematisch die Methodik der statistischen Bearbeitung eines für anthropometrische Bevölkerungsvergleiche in Graubünden (Schweizer Alpen) aufgenommenen Zahlenmaterials. Zwei in ihrer Sprache, Abstammung und Ethnologie unterscheidbare Klassen Einwohner werden durch acht Körpermasse voneinandergetrennt, nämlich: Stammhöhe, Beinlänge (= Differenz zwischen Körperhöhe und Stammhöhe), grösste Kopfbreite und -Länge, Jochbogenbreite, morphologische Gesichtshöhe, Breite und Höhe der Nase.

Aus den Vergleichstests folgt insgesamt, dass sich die Eigenart der beiden Teilbevölkerungen anthropologisch bei Männern und Frauen ziemlich ähnlich und jedenfalls durch eine bei beiden Geschlechtern statistisch gesicherte Trennung äussert.

Anderseits ergibt sich aus den Vereinfachungstests, dass für diese Trennung Stammhöhe, Beinlänge, grösste Kopfbreite und Nasenlänge den wesentlichsten Ausschlag geben.

Diese zwei Ergebnisse rechtfertigen gemeinsam die Bestimmung einer einzigen von diesen vier Massen abhängigen und für beide Geschlechter gleichzeitig gültige Trennformel, die ihrerseits ohne merklichen Informationsverlust durch eine noch einfachere lineare Formel mit ganzzahligen Koeffizienten angenähert werden kann.

Die für das gegenwärtige Trennverfahren massgebende Bedeutung der genannten vier Merkmale kann auf die verschiedenen gegenseitigen Merkmalsabhängigkeiten zurückgeführt werden und wird tatsächlich durch den numerischen Wert der einzelnen einfachen Korrelationen zwischen je zwei Merkmalen durchaus bestätigt.

Der Ermittlung der Einzelwertverteilung sämtlicher am Trennverfahren mitwirkenden Teilbevölkerungen geht die Berechnung des Wertes der vereinfachten, vom Geschlecht unabhängigen Trennfunktion für jeden untersuchten Bewohner voraus. Auf diese Einzelwertverteilungen stützt sich die Untersuchung über die gegenseitigen Lagebeziehungen der verschiedenen Teilbevölkerungen. Ebenfalls geht das Anwendungsbeispiel der Zuordnung einer am Trennverfahren nicht beteiligten Einwohnergruppe von diesen Einzelwertverteilungen aus.

3.2 LISTE DES SYMBOLES ET ABRÉVIATIONS

La référence placée entre parenthèses à la suite de la définition indique le paragraphe où celle-ci est introduite, ou énoncée pour la première fois.

$A_{ij(p)}$	régression linéaire multiple de x_{ijp} en $x_{ij1}, x_{ij2}, x_{ij3}, \dots, x_{ijp-1}$ (§ 1.54).
B	coefficient de détermination d'une corrélation multiple (§ 1.12).
$B(x_k; x_l)$	coefficient de détermination de la corrélation simple entre x_k et x_l (§ 2.43).
C	constante d'intégration (§ 1.11).
$Cov(\dots)$	covariance totale, sans égard à aucune classification (§ 1.12).
CRL	coefficient de ressemblance raciale de K. PEARSON (§ 1.14).
D, D'	anciennes expressions de la distance entre deux échantillons (§ 1.14).
D^2, D_p^2, D_{p+q}^2	distance généralisée de MAHALANOBIS, basée sur p ou bien $p + q$ variables (caractères) (§ 1.13, § 1.51).
D_G^2, D_H^2	distances associées à des fonctions H ou bien G (§ 1.41, § 2.42).
Da	localité de <i>Davos</i> (fig. 1 et 2, § 2.22).
DL	degrés de liberté (tabl. 8, § 1.41).
E	espérance mathématique (§ 1.11).
$E(x_{ijk})$	espérance mathématique par rapport à j, i et k étant constants (§ 1.32).
$F^{(j)}(f_1; f_2)$	répartition de F avec degrés de liberté f_1 et f_2 (§ 1.12).
F_P	limite de signification de la répartition F au seuil P (§ 2.43).
Fu	localité de <i>Furna</i> (fig. 1 et 2, § 2.22).
$G; G_{ij}$	fonction discriminante empirique à coefficients désignés d'avance; valeurs individuelles de cette fonction (§ 1.41).
$\bar{G}, \bar{G}_i$	moyenne générale et moyenne de l'échantillon i des valeurs de la fonction G (§ 1.41).
G_p	fonction empirique basée sur p caractères (variables) (§ 2.42).
Gl	localité de <i>Glas</i> (fig. 1 et 2, § 2.22).
H, H_p, H_{ij}	meilleure fonction discriminante linéaire, basée sur p variables (caractères); valeurs individuelles de cette fonction (§ 1.11).
$\bar{H}, \bar{H}_i$	moyenne générale, moyenne de l'échantillon i des valeurs de la fonction H (§ 1.11).
$\bar{H}', \bar{H}''$	moyenne des valeurs H pour divers sous-ensembles (§ 2.53).
$(\bar{H}' - \bar{H}'')_P$	plus petite différence significative, au seuil P , de deux moyennes de la fonction H (§ 2.53).
I, II	signes distinctifs des deux types (type I, type II) (§ 2.11).
J_1, J_2	mesures de non-normalité (obliquité, <i>kurtosis</i>) (§ 2.54).
$K_1, K_2, \text{etc.}$	estimations des cumulants successifs d'une répartition (§ 2.53).
Kl	localité de <i>Klosters</i> (fig. 1 et 2, § 2.22).
L_n	logarithme naturel, base e (§ 1.31).
Lw	localité de <i>Langwies</i> (fig. 1 et 2, § 2.22).
N, N_i	effectif total, effectif de l'échantillon i (§ 1.11).
$N_{WIO}, N_{RII\varnothing}, \text{etc.}$	effectif des sujets masculins walsers du type I, ou des sujets féminins romanches du type II, etc. (§ 2.53).
N', N''	effectif des sujets dans divers sous-ensembles (§ 2.53).
$[0]$	matrice nulle (§ 1.21).
P	probabilité, seuil de probabilité ou de signification (§ 1.12).
R, RI, RII	Romanches, sujets romanches du type I ou II (§ 2.11).
$S_{kl}, S'_{kl}, S''_{kl}$	somme des produits corrigée totale, entre les échantillons, ou à l'intérieur des échantillons, pour les variables (ou caractères) x_k et x_l (§ 1.12).
S_H, S'_H, S''_H	sommes des carrés corrigée totale, entre les échantillons ou à l'intérieur des échantillons, des valeurs H_{ij} (§ 1.12).
S_{yy}	somme des carrés corrigée totale des valeurs y_{ij} (§ 1.12).

S_{ky}	somme des produits corrigée totale des variables x_k et y (§ 1.12).
S_r	somme des carrés résiduelle dans une régression (§ 1.54).
Sal	localité de <i>Salouf</i> (fig. 1 et 2, § 2.22).
Sav	localité de <i>Savognin</i> (fig. 1 et 2, § 2.22).
St.A	localité de <i>St. Antönien</i> (fig. 1 et 2, § 2.22).
T^2	rapport généralisé de HOTELLING (notation originale) (§ 1.14).
T_p, T_{p+q}	rapport généralisé de HOTELLING (notation de RAO), basé sur p ou bien $p + q$ variables (caractères) (§ 1.13).
Ti	localité de <i>Tinizong</i> (fig. 1 et 2, § 2.22).
$Tsch$	localité de <i>Tschappina</i> (fig. 1 et 2, § 2.22).
U	variable auxiliaire sans signification propre (§ 1.42).
$Var (...)$	variance totale, sans égard à aucune classification (§ 1.12).
W, WI, WII	Walsers, sujets walsers du type I ou II (§ 2.11).
Wie	localité de <i>Wiesen</i> (fig. 1 et 2, § 2.22).
X_{ijk}	résultat (brut) de la mesure du k -ième caractère sur le sujet numéro j de l'échantillon numéro i (§ 2.11).
$\bar{Y}, \bar{Y}_{ij}$	régression linéaire multiple de y_{ij} en x_{ij} ($k = 1, ..., p$) (§ 1.12).
$\bar{Y}, \bar{Y}_i$	moyenne générale, moyenne de l'échantillon i des valeurs de régression Y_{ij} (§ 1.41).
Z_k	constante de transformation linéaire (§ 2.11 et tabl. 27, § 3.5).
a_{kl}	élément général de l'estimation de la matrice $[\alpha_{kl}]$ (§ 1.13).
a^{kl}	élément général de la matrice inverse $[\alpha_{kl}]^{-1}$ (§ 1.13).
$adj [M_{kl}]$	matrice adjointe de la matrice $[M_{kl}]$ (§ 2.42).
b_k	coefficient de régression linéaire multiple (§ 1.12).
$b_{y.G}, b_{Y.G}$	coefficient de régression linéaire simple de y_{ij} ou Y_{ij} en G_{ij} (§ 1.41).
c^2	constante égale, par convention, à $\sum_{ij} G_{ij}^2$ (§ 1.41).
$cov (...)$	covariance à l'intérieur des échantillons (§ 1.12).
chap.	abréviation pour « chapitre ».
d_k	k -ième variable d'une répartition normale p -dimensionnelle (§ 1.13); écart des moyennes d'échantillon de la variable (caractère) k -ième (§ 1.21).
d'_k, d''_k	écarts, pour la k -ième variable (caractère), des moyennes de deux échantillons i' et i'' (§ 1.32).
d_H	écart des moyennes d'échantillon des valeurs H_{ij} (§ 1.21).
d'_k, d'_{1k}, d'_{2k} $d_{1,k}, d_{2,k}, ...$	} quantités intermédiaires dans un procédé d'élimination (tabl. 12, § 1.62).
dx_k	
e	élément différentiel de la variable x_k (§ 1.11).
e, e'_k	constante numérique, base des logarithmes naturels (§ 1.11).
e_k, e'_k	combinaisons linéaires des écarts d_k (§§ 1.32 et 1.43).
eq.	abréviation pour « équation ».
$exp (...)$	puissance de e dont l'exposant est l'argument (...) (§ 1.11).
f, f_1, f_2	degrés de liberté (§ 1.12).
form.	abréviation pour « formule ».
g_k	coefficients d'une fonction linéaire empirique, fixés d'avance (§ 1.41).
$h_k, h_{p,k}$	coefficients de la meilleure fonction discriminante linéaire calculée, basée sur p variables (caractères) (§§ 1.11 et 1.54).
h'_k	système de coefficients proportionnels aux h_k (§ 2.1).
i, i', i''	indices relatifs aux échantillons ($i = 1, ..., m$) (§ 1.11).
j	indice relatif à l'individu dans l'échantillon ($j = 1, ..., N_i$) (§ 1.11).
k	} indices relatifs à la variable (au caractère); $k = 1, ..., p$ ou bien $p + q$ (§ 1.11).
l	
m	
	nombre d'échantillons (de sous-populations) (§ 1.11).

n, n_i	degrés de liberté d'une estimation (§ 1.11; § 2.53).
p	nombre de variables considérées (caractère; dimension) (§ 1.11).
q	nombre de variables (caractères) supplémentaires considérées outre les p précédentes (§§ 1.13 et 1.51).
$r, r(x_k; x_l)$	coefficient de corrélation simple totale (corrélation de produit des moments) entre les variables x_k et x_l sans égard à aucune subdivision (§ 1.12).
$r'', r''(x_k; x_l)$	coefficient de corrélation simple, entre x_k et x_l , à l'intérieur des échantillons (§ 1.12).
r_p	coefficient de corrélation multiple (p variables indépendantes) (§ 1.12).
r_d^2	rapport de deux sommes de carrés égal à $r_p^2 (1 - r_p^2)$ (§ 1.41).
s_k	écart-type (déviations-standard) de la variable x_k ou du k -ième caractère (§§ 1.14 et 2.12).
s^2	variance à l'intérieur des échantillons (§ 1.13).
s_{kl}	dispersion estimée des variables x_k, x_l à l'intérieur des échantillons (§§ 1.14, 1.51).
s^{kl}	élément général de l'inverse $[s_{kl}]^{-1}$ (§ 1.14).
$s_{H,i}^2, s_{H,i}^2$	variance des valeurs H_{ij} ; variance à l'intérieur des échantillons (§ 2.53).
s_J	déviations-standard (écart-type) de la mesure de non-normalité J (§ 2.53).
$t_{[f]}$	répartition de t (« STUDENT ») à f degrés de liberté (§ 1.13).
tabl.	abréviation pour « tableau ».
u	variable normale standardisée (§ 2.53).
var	variance à l'intérieur des échantillons (§ 1.12).
w_{kl}	dispersion estimée des variables (caractères) x_k et x_l à l'intérieur des échantillons (§ 1.13).
w^{kl}	élément général de l'inverse $[w_{kl}]^{-1}$ (§ 1.13).
w_{ekl}	dispersion estimée des quantités e_k, e_l (ou bien des quantités e'_k, e'_l) (§ 1.32).
w_p^{kl}	élément général de la matrice inverse $[w_{kl}]^{-1}$ lorsque $k = 1, \dots, p$ et $l = 1, \dots, p$ (§ 1.54).
$w_{kl}^1, w_{kl}^2, \dots$	quantités intermédiaires dans un procédé d'élimination (tabl. 11, § 1.61).
$w_{k\tau}, w_{k\tau}^1, w_{k\tau}^2$	totaux de contrôle dans un procédé d'élimination (tabl. 11, § 1.61).
$w_{1,kl}, w_{2,kl}$	} quantités intermédiaires dans un procédé d'élimination (tabl. 12, § 1.62).
$w'_{1,kl}, w'_{2,kl}$	
$w_{1,k\tau}, w_{2,k\tau}$	} totaux de contrôle dans un procédé d'élimination (tabl. 13, § 1.63).
$w_{1,2\tau}, w_{2,3\tau}$	
$w^{k\tau}$	
x_k	k -ième variable dans une répartition normale p -dimensionnelle (§ 1.11).
x_{ijk}	valeur transformée de la mesure k sur l'individu $(i; j)$ (§ 2.11); variable indépendante dans une régression linéaire multiple (§ 1.12).
$\bar{x}_{ij}^\tau$	sous-total de contrôle (§ 2.12).
$\bar{x}_k, \bar{x}_{ik}$	moyenne générale, moyenne de l'échantillon i de la variable (du caractère) k -ième (§ 1.11).
$\bar{x}, \bar{x}'$	moyenne, dans deux échantillons différents, d'un unique caractère mesurable (§ 1.13).
y, y_{ij}	variable dépendante d'une régression linéaire (simple ou multiple), pseudo-variable ($= \pm N_i/N$) dans une régression (§ 1.22).
$\bar{y}, \bar{y}_i$	moyenne générale, moyenne de l'échantillon i des valeurs y_{ij} (§ 1.22).
z	répartition de z de R. A. FISHER (§ 1.14).
Γ, Γ_{ij}	régression linéaire simple de Y_{ij} en G_{ij} (§ 1.41).
Δ_p^2	distance généralisée dans l'ensemble de base (§ 1.13).
$\sum, \sum \sum$	sommation, double sommation (§ 1.11).

$\int_{\Omega}$	intégration (§ 1.11). forme quadratique à matrice $ \omega_{kl} $ (§ 1.11).
α_{kl}	élément général de la matrice de dispersion d'une répartition normale p -dimensionnelle de variables d_k (§ 1.13).
β_{kl} $[\beta^{kl}]$	élément général d'une matrice dépendant des quantités σ_{ik} et ρ_{ikl} (§ 1.14). matrice inverse $[\beta_{kl}]^{-1}$ (§ 1.14).
γ_{ij}	régression linéaire simple de y_{ij} en G_{ij} (§ 1.41).
δ_{kl}	symbole de KRONECKER (§§ 1.54 et 1.63).
∂_k	dérivation partielle par rapport à la variable x_k (§ 1.21).
ε	constante auxiliaire égale à $\eta(1 - r_p^2)$ (§ 1.22).
η	constante numérique égale à $N_1 N_2 / (N_1 + N_2)$, demi-moyenne harmonique (§ 1.13).
θ (h_k)	fonction des coefficients h_k , rapport de deux sommes de carrés (§ 1.21).
θ'	variable auxiliaire égale à $\theta/(1 + \theta)$ (§ 1.21).
λ	constante de proportionnalité non nulle (§ 1.21).
μ_k	espérance mathématique de la variable x_k d'une répartition normale p -dimensionnelle (paramètre de position) (§ 1.11).
ν	constante numérique égale généralement à η/n ($n = N_1 + N_2 - 2$) (§ 1.13).
ζ_k	variable auxiliaire égale à $d_k \sqrt{\eta}$ (§ 1.31).
ρ_{ikl}	corrélation des variables (caractères) x_k, x_l dans l'ensemble de base i -ème (§ 1.14).
σ_{ik}	déviations-standard (écart-type) de la variable (caractère) x_k dans l'ensemble de base i -ème (§ 1.14).
τ	indice spécifique d'un sous-total de contrôle (§§ 1.61 et 2.12).
φ	densité de probabilité de la répartition des variables x_k (§ 1.11).
$\chi_{[f]}^2$	répartition de chi-carré à f degrés de liberté (§ 2.53).
ψ	fonction numérique des effectifs (le plus souvent $\{\sum_i (1/N_i)\}^{-1}$) (§ 1.32).
ω_{kl}	élément général de la matrice de dispersion de l'ensemble de base, ou de la matrice de la forme quadratique d'une répartition normale p -dimensionnelle (§ 1.11).
$[\omega^{kl}]$	inverse de la matrice $[\omega_{kl}]$ (§ 1.11).
$[M_{kl}]$	matrice d'élément général M_{kl} (§ 1.11).
$[M^{kl}], [M_{kl}]^{-1}$	inverse de la matrice $[M_{kl}]$; élément général: M^{kl} (§ 1.11).
$ M_{kl} $	déterminant de la matrice $[M_{kl}]$ (§ 2.42).
$ M_{kl} ^{-1}$	déterminant de $[M_{kl}]^{-1}$, inverse du déterminant $ M_{kl} $ (§ 2.42).
$[h_1, \dots, h_p]$	vecteur-ligne ou vecteur-colonne d'éléments h_k (§§ 1.21 et 1.61).
$ \dots $	valeur absolue (§ 1.13).
$\♂, \text{♀}$	hommes, femmes (masculin, féminin) (§ 2.21).
$\circ$	en notation exponentielle: résultat non significatif d'un test statistique (au seuil de 5%) (§ 1.12).
$*, **$	résultat significatif (seuil $P = 0.05$) ou hautement significatif ($P = 0.01$) d'un test statistique (§ 1.12).

Numérotation des variables (caractères):

- | | |
|---|---|
| 1 | buste (taille assis) |
| 2 | longueur des jambes |
| 3 | diamètre antéro-postérieur maximum du crâne |
| 4 | diamètre transverse maximum du crâne |
| 5 | diamètre bizygomatique du crâne |
| 6 | hauteur faciale (morphologique) |
| 7 | hauteur du nez |
| 8 | largeur du nez (cf. § 2.11 et tabl. 27, § 3.5). |

En cas d'incertitude quant à la terminologie statistique française, nous nous sommes assez généralement conformé aux indications du « Dictionnaire de termes statistiques » de KENDALL and BUCKLAND (*A dictionary of statistical terms*, 1957), de même qu'aux deux ouvrages de LAMOTTE (1948, 1957).

3.3. REMERCIEMENTS

C'est un plaisir pour nous que d'exprimer à cet endroit notre profonde gratitude envers le directeur du Laboratoire de Statistique mathématique de l'Université de Genève, M. le professeur A. LINDER, à l'enseignement et la vaste expérience de qui nous sommes redevable de l'essentiel de notre formation. Il est hors de doute que sans l'aide de son intérêt soutenu et de ses directives perspicaces quant aux sujets mêmes à tirer au clair dans cette analyse anthropométrique, nous ne serions guère parvenu au terme de la présente thèse.

Nos remerciements s'adressent également au directeur de l'Institut d'Anthropologie de l'Université de Genève, M. le professeur M.-R. SAUTER, et à son chef de travaux, M^{lle} H. KAUFMANN, D^r ès sc., ainsi qu'à M. K. HAEGLER, D^r ès sc. (ancien directeur du Musée d'histoire naturelle à Coire), pour la patience et l'effort de compréhension toujours à nouveau manifestés à l'égard du non-anthropologiste désireux de mettre à l'épreuve des concepts ou des techniques nouveaux. Les échanges de vues qui en ont résulté ont été pour nous, est-il besoin de le dire, la source d'un constant enrichissement.

M^{lle} H. KAUFMANN et M. K. HAEGLER, tous deux anthropologistes de la « Communauté de recherches séro-anthropologiques et génétiques aux Grisons », ont aimablement mis à notre libre disposition l'ensemble des données métriques recueillies par eux au cours d'une enquête de ladite communauté réalisée en 1954 avec l'aide du Fonds national suisse de la recherche scientifique; qu'ils en soient vivement remerciés.

A M^{lle} M. SCHNEEBERGER, assistante, nous sommes reconnaissant pour la part importante qu'elle a prise dans la préparation du manuscrit, l'exécution de toutes les figures en particulier.

Enfin, nous adressons nos plus vifs remerciements à la Commission de la Bourse J. et N. LEBEDINSKY et au Fonds KRISHNAVARMA pour l'aide financière apportée à ce travail.

3.4 BIBLIOGRAPHIE

- AITKEN, T. W.: *Determinants and matrices*. Edinburgh and London, Oliver & Boyd, 6th ed. (1949).
- ANDERSON, T. W.: *Introduction to multivariate statistical analysis*. New York, Wiley & Sons/ London, Chapman & Hall (1958).
- BARLOW, P.: *Tables of squares, cubes, square roots, cube roots and reciprocals*. London, Spon, 4th ed. repr. (1947).
- BARNARD, M. M.: *The secular variation of skull characters in four series of Egyptian skulls*. Ann. Eugenics 6, pt. 4 (1934-35), 352.
- BARTLETT, M. S.: *Some examples of statistical methods of research in agricultural and applied biology*. J. Roy. Statist. Soc. Suppl. 4, n° 2 (1937), 137.
- *Further aspects of the theory of multiple regression*. Proc. Cambridge Phil. Soc. 34 (1938), 33.
- *The standard errors of discriminant function coefficients*. J. Roy. Statist. Soc. Suppl. 6, n° 2 (1939a), 169.
- *A note on tests of significance in multivariate analysis*. Proc. Cambridge Phil. Soc. 35 (1939b), 180.

- BONÉ, E. L.: *L'influence de la hauteur du buste sur l'allométrie des segments particuliers chez l'homme et divers autres primates*. Arch. suisses d'Anthropologie générale, 18 (1953), 1.
- BORTH, R.: *Directives pour la préparation et l'impression de manuscrits*. Acta Endocrinologica (1959).
- BOSE, R. C. and ROY, S. N.: *The distribution of the studentised D^2 -statistic*. Sankhyā, prt. 1 (1938), 19.
- BOSE, P. K.: *On recursion formulae, tables and BESSEL function populations associated with the distribution of classical D^2 -statistic*. Sankhyā 8, part. 3 (1947), 235.
- CHOWDHURY, S. B.: *The most powerful unbiased critical regions and the shortest unbiased confidence interval associated with the distribution of classical D^2 -statistic*. Sankhyā 14, prt. 1 and 2 (1954), 71.
- COCHRAN, W. G.: *The omission or addition of an independent variate in multiple linear regression*. J. Roy. Statist. Soc. Suppl. 5, n° 2 (1938), 171.
- DWYER, P. S.: *Linear computations*. New York, Wiley & Sons/London, Chapman & Hall (1951).
- EAST, D. A. and OSCHINSKY, L.: *A comparison of serological and somatometrical methods used in differentiating between certain East African racial groups, with special reference to D^2 -analysis*. Sankhyā 20, prt. 1 (1958), 31.
- ELDERTON, E. M. and WOO, T. L.: *On the normality or want of normality in the frequency distributions of cranial measurements*. Biometrika 24 (1932), 45.
- EZEKIEL, M. and FOX, K. A.: *Methods of correlation and regression analysis, linear and curvilinear*. New York, Wiley & Sons/London, Chapman & Hall, 3rd ed. (1959).
- FISHER, R. A.: *The use of multiple measurements in taxonomic problems*. Ann. Eugenics 7, pt. 2 (1936-37), 179.
- *The statistical utilization of multiple measurements*. Ann. Eugenics 8, pt. 4 (1937-38), 376.
- *The precision of discriminant functions*. Ann. Eugenics 10, pt. 4 (1940), 422.
- *Statistics*. Sankhyā 12, prt. 3 (1953), 303.
- *Statistical methods for research workers*. Edinburgh and London, Oliver & Boyd, 13th ed. (1958).
- HALD, A.: *Statistical theory with engineering applications*. New York, Wiley & Sons/London, Chapman & Hall (1952a).
- *Statistical tables and formulae*. New York, Wiley & Sons/London, Chapman & Hall (1952b).
- HOTELLING, H.: *The generalization of STUDENT's ratio*. Ann. math. Statistics 2 (1931), 360.
- HOTZENKOECHERLE, R.: *Zur Sprachgeographie Deutschbündens mit besonderer Berücksichtigung des Verhältnisses zum Wallis*. Jahresber. Hist.-ant. Ges. Graubünden 74 (1944), 135.
- KAUFMANN, H., HAEGLER, K., LANG, R.: *Analyse anthropologique et statistique de Walsers orientaux et de Romanches de l'Oberhalbstein*. Arch. suisses d'Anthropologie générale 1-2 (1958), 1.
- KENDALL, M. G.: *The advanced theory of statistics*. Vol. 2. London, Griffin (1946).
- and BUCKLAND, W. G.: *A dictionary of statistical terms*. Edinburgh and London, Oliver & Boyd (1957).
- LAMOTTE, M.: *Introduction à la biologie quantitative*. Paris, Masson (1948).
- *Initiation aux méthodes statistiques en biologie*. Paris, Masson (1957).
- LETESTU, S.: *Note sur l'analyse discriminatoire*. Experientia, Vol. IV/1 (1948), 22.

- LINDER, A.: *Statistische Methoden für Naturwissenschaftler, Mediziner und Ingenieure*. Basel, Birkhäuser, 2. erw. Aufl. (1957).
- *Statistische Methoden für Naturwissenschaftler, Mediziner und Ingenieure*. Basel, Birkhäuser, 3. erw. Aufl. (1960).
- *Planen und Auswerten von Versuchen*. Basel, Birkhäuser (1953).
- MAHALANOBIS, P. C.: *Analysis of race mixture in Bengal*. J. and Proc. Asiat. Soc. of Bengal, New Series 23 (1927), 301.
- *On tests and measures of group divergence. Part 1: Theoretical formulae*. J. and Proc. Asiat. Soc. of Bengal, New Series 26 (1930), 541.
- *Historical note on the D^2 -statistic*. In: MAHALANOBIS, MAJUMDAR and RAO (1949), Appendix 1.
- MAJUMDAR, D. N., RAO, C. R.: *Anthropometric survey of the United Provinces 1941. A statistical study*. Sankhyā 9, prt. 2-3 (1949), 89.
- and RAO, C. R.: *Statistical analysis*. In: MAHALANOBIS, MAJUMDAR and RAO (1949), Part II.
- MUKHERJEE, R.: *A study on differences in physical developments by socio-economic strata*. Sankhyā 11, prt. 1 (1951), 47.
- PEARSON, K.: *On the coefficient of racial likeness*. Biometrika 28 (1926), 105.
- PILLAI, K. C. S.: *Concise tables for statisticians*. Manila, The Statistical Center, University of the Philippines (1957).
- PIQUET, M. M.: *Essai de standardisation de la mesure de la taille assis*. Bull. Mém. Soc. Anthropol. Paris 5, 10^e série (1954), 57.
- QUENOUILLE, M. H.: *Note on the elimination of insignificant variates in discriminatory analysis*. Ann. Eugenics 14, pt. 4 (1947-49), 305.
- RAO, R. C.: *Tests with discriminant functions in multivariate analysis*. Sankhyā, 7, prt. 4 (1946), 407.
- *Advanced statistical methods in biological research*. New York, Wiley & Sons/London, Chapman & Hall (1952).
- SNEDECOR, G. W.: *Statistical methods*. Ames, Iowa State College Press, 5th ed. (1956).
- STEVENS, W. L.: *Análise discriminante. Questões de método*, 13. Rev. da Fac. de Ciencias da Univ. de Coimbra 23, n° 1 (1945), 1.
- STIEFEL, D.: *Lehrbuch der darstellenden Geometrie*. Basel, Birkhäuser (1947).
- WALD, A.: *On a statistical problem arising in the classification of an individual into one of two groups*. Ann. math. Statistics 15 (1944), 145.
- WILLIAMS, E. J.: *Regression analysis*. New York, Wiley & Sons/London, Chapman & Hall (1959).
- WINICK, Ch.: *Dictionary of anthropology*. London, Peter Owen Ltd. (1957).
- WOOLF, B.: *Computation and interpretation of multiple regressions*. J. Roy. Statist. Soc. 13 (1951), 100.

A titre documentaire, nous donnons ci-dessous une liste de quelques travaux — non cités dans notre exposé — qui s'apparentent à nos recherches (applications de l'analyse multidimensionnelle à des données anthropométriques sur le vivant). Cette liste ne prétend pas être exhaustive.

- BAUER, K.: *Methoden der simultanen Analyse mehrerer anthropologischer Merkmale*. Ber. über die 5. Tagung der Deutschen Ges. f. Anthropologie 1956, Göttingen, Muster-schmidt-Verlag (1957), 171.

- BHATTACHARYYA, A.: *On a measure of divergence between two multinomial populations.* Sankhyā 7, prt. 4 (1946), 401.
- DIAZ UNGRIA, A. D., CAMACHO, A., RIOS, S.: *Analisis multivariante de dos muestras de indigenas de Venezuela: Caribe y Guarao.* Boletin Indigenista Venezolano. Años III, IV y V — Tomos III, IV y V, Caracas (1955, 1956 y 1957) — Nos. 1-4 (Marzo de 1958).
- JOHNSON, O. P.: *The quantification of qualitative data in discriminant analysis.* J. Amer. Statist. Assoc. 45 (1950), 65.
- LINDER, A., WEBER, A. A., MORGENTHALER, P. W.: *Application de l'analyse discriminante à l'étude des différences anthropométriques selon les groupes sanguins du système ABO.* Experientia XIII/3 (1957), 110.
- MAJUMDAR, D. N. and RAO, C. R.: *Bengal anthropometric survey 1945: A statistical study.* Sankhyā 19, prt. 3 and 4 (1958), 201.
- MAUNG, K.: *Discriminant analysis of TOCHER's eye colour data for scottish school children.* Ann. Eugenics 11, pt. 1 (1941-42), 64.
- RAO, C. R.: *The utilization of multiple measurements in problems of biological classification.* J. Roy. Statist. Soc. B 10, No. 2 (1948), 159.
- SCHWIDETZKY, I.: *Morphologie und Statistik im Vaterschaftsnachweis.* Homo 6 (1955), 46.
- SINISCALCO, M., MONTALENTI, G., SILVESTRONI, E., BIANCO, I.: *Effects of a single gene difference on the pattern of some physical measurements.* In: A CIBA Foundation on Medical Biology and Etruscan Origins (Ed. by G. E. W. WESTENHOLM and C. M. O'CONNOR). London, Churchill (1959), 205.
- SMITH, S. M.: *Discrimination between electro-encephalograph recordings of normal females and normal males.* Ann. Eugenics 18, pt. 4 (1953-54), 344.
- TRAVERS, R. M. W.: *The use of a discriminant function in the treatment of psychological group differences.* Psychometrika 4 (1939), 25.
- WALLACE, N. and TRAVERS, R. N. W.: *A psychometric sociological study of a group of speciality salesmen.* Ann. Eugenics 8, pt. 3 (1937-38), 266.
- WEBER, A. A.: *Efficacité de l'indice cranien et de l'indice nasal comparée à l'analyse discriminante.* Acta Genetica et Statistica Medica II/4 (1951), 351.

3.5 ANNEXE

Dans cette annexe sont reproduits un certain nombre de tableaux contenant des données numériques et des résultats intermédiaires, tableaux qui pour ne pas être indispensables pour la compréhension de l'exposé pourraient cependant se trouver utilisables dans d'éventuelles recherches ultérieures.

TABLEAU 27

Numérotation des caractères (k), constantes Z_k de changement d'échelle, sommes partielles $\sum_j x_{ijk}$ de huit caractères et effectif N des observations dans les sous-populations principales

			Walsers (W)				Romanches (R)			
			Hommes		Femmes		Hommes		Femmes	
Caractères	k	Z_k	I	II	I	II	I	II	I	II
Buste	1	50	5843	4528	4396	4747	2371	1118	1931	330
Longueur des jambes .	2	50	4900	3727	3158	3435	1855	886	1331	245
Diamètre antéro-postér.	3	150	6460	4826	4057	4366	2555	1125	1778	328
Diamètre transverse . .	4	100	8673	6515	6280	6664	3186	1507	2482	463
Diamètre bizygomatique	5	100	6268	4725	4138	4425	2387	1130	1672	314
Hauteur faciale	6	90	4874	3669	2597	2959	2037	913	1243	181
Hauteur du nez	7	0	8024	6285	6146	6575	3316	1494	2742	498
Largeur du nez	8	0	5399	4144	4133	4378	2077	1003	1763	320
Effectif (N)			151	117	126	134	60	28	54	10

Sommes $\sum_j x_{ijk} x_{ijl}$ de produits non corrigées ($k = 1, \dots, 8$; $l = 1, \dots, 8$);
sous-populations walsers

Walsers I ♂									
	1	2	3	4	5	6	7	8	k/l
	227841	190711	250596	336015	243092	189447	310406	209021	1
		161672	210430	281875	204043	159096	260470	175500	2
1	176484		281892	371754	268785	208738	342904	231449	3
2	144611	120819		502973	362707	280501	461184	310519	4
3	187004	153771	203944		264310	203104	333353	224609	5
4	252466	207540	269837	366339		162760	260611	174271	6
5	183325	151216	196304	265319	194621		428378	287066	7
6	142437	117093	152448	204882	148898	119833		194125	8
7	243363	200475	259757	350505	254377	198477	339155		
8	160294	132037	171482	231279	167999	130063	222896	147584	
l/k	1	2	3	4	5	6	7	8	
Walsers II ♂									
Walsers I ♀									
	1	2	3	4	5	6	7	8	k/l
	154380	110275	142072	219316	144612	90980	214460	144203	1
		81214	101632	157122	103652	65685	154169	103452	2
1	169307		134781	202940	134155	84585	198137	133565	3
2	121833	90099		316796	208632	129803	306528	206589	4
3	155699	112799	147356		139170	86028	202148	136213	5
4	236245	170943	217595	335060		57971	128060	85081	5
5	157147	113559	145071	220011	149057		301130	201661	7
6	105275	76479	97954	148051	98370	69557		136491	8
7	233076	168635	214652	327565	217506	145989	323823		
8	155187	112249	143058	217959	144953	96774	214919	143834	
l/k	1	2	3	4	5	6	7	8	
Walsers II ♀									

TABLEAU 30

$$\text{Dispersion } w_{kl} = \sum_{i,j} (x_{ijk} - \bar{x}_{ik}) (x_{ijl} - \bar{x}_{il}) / (N - 2)$$

pour les discriminations typiques

Discrimination WI♂/WII♂									
	1	2	3	4	5	6	7	8	<i>k/l</i>
	11.2425	5.5512	3.2241	2.7821	3.8080	4.8472	0.1599	0.0842	1
		17.8267	3.1619	1.6538	5.0634	4.3278	1.3412	1.2473	2
1	8.3380		39.1188	6.8331	7.6642	5.0029	0.5254	3.8450	3
2	0.9409	15.9247		31.5054	18.4382	4.2516	3.1631	3.5401	4
3	6.0450	3.2129	35.8707		29.8100	5.6821	3.1466	4.2933	5
4	1.4887	−0.6236	4.6586	28.8494		38.3956	11.2669	0.4237	6
5	2.4477	0.2572	7.0280	9.0645	24.0533		13.2623	1.7179	7
6	3.1972	4.7370	9.7260	4.8883	5.4108	33.5660		7.1143	8
7	0.7250	0.8452	2.5981	3.0425	2.6719	8.4627	9.8766		
8	0.3951	−0.4400	3.4983	3.2179	3.3373	−0.0237	0.6411	6.6651	
<i>l/k</i>	1	2	3	4	5	6	7	8	
Discrimination WI ♀/WII ♀									
Discrimination RI ♂/RII ♂									
	1	2	3	4	5	6	7	8	<i>k/l</i>
	10.4749	7.3589	10.6660	4.4598	5.5166	6.8228	1.6680	−0.0171	1
		30.2676	17.0663	3.6013	6.7266	17.1840	2.9989	−0.4152	2
1	8.4818		77.7755	17.5860	10.6953	22.1362	9.4754	0.2827	3
2	−0.0092	18.7228		41.6222	24.8963	5.3992	1.4324	3.9117	4
3	5.7587	−2.6033	29.5049		32.5885	5.0265	−0.5477	4.9163	5
4	3.8148	−3.9226	15.4105	28.6456		53.5992	15.9620	−0.9199	6
5	3.8632	−0.8984	8.3008	15.8020	24.8440		14.6530	0.3014	7
6	4.6652	7.9492	14.1173	11.7377	12.4296	40.9336		6.7708	8
7	1.1308	2.0287	2.9308	5.5993	4.7315	11.3778	15.1441		
8	1.1843	−0.6078	3.4934	0.6983	3.7467	4.1186	2.4319	4.7793	
<i>l/k</i>	1	2	3	4	5	6	7	8	
Discrimination RI ♀/RII ♀									

TABLEAU 33

Matrices $[(d'_k \pm d''_k) (d'_l \pm d''_l)]$ pour la comparaison des fonctions discriminantes raciales complètes

[(d'_k + d''_k) (d'_l + d''_l)]									
	1	2	3	4	5	6	7	8	k/l
1 2 3 4 5 6 7 8	2.8629	-3.2960	-1.5668	-13.8998	-6.0997	6.9034	6.9846	-2.1860	1
		3.7947	1.8038	16.0028	7.0225	-7.9478	-8.0413	2.5168	2
	0.0023		0.8575	7.6071	3.3382	-3.7781	-3.8225	1.1964	3
	0.0537	1.2499		67.4862	29.6151	-33.5172	-33.9115	10.6138	4
	-0.0254	-0.5925	0.2809		12.9960	-14.7084	-14.8814	4.6577	5
	0.0220	0.5131	-0.2433	0.2107		16.6464	16.8422	-5.2714	6
	-0.0072	-0.1688	0.0800	-0.0693	0.0228		17.0404	-5.3334	7
	0.0353	0.8228	-0.3901	0.3378	-0.1111	0.5417		1.6693	8
7	-0.0061	-0.1431	0.0678	-0.0588	0.0193	-0.0942	0.0164		
8	0.0472	1.1001	-0.5215	0.4516	-0.1486	0.7242	-0.1260	0.9682	

k/l	1	2	3	4	5	6	7	8
1	2.8629	-3.2960	-1.5668	-13.8998	-6.0997	6.9034	6.9846	-2.1860
2		3.7947	1.8038	16.0028	7.0225	-7.9478	-8.0413	2.5168
3	0.0023		0.8575	7.6071	3.3382	-3.7781	-3.8225	1.1964
4	0.0537	1.2499		67.4862	29.6151	-33.5172	-33.9115	10.6138
5	-0.0254	-0.5925	0.2809		12.9960	-14.7084	-14.8814	4.6577
6	0.0220	0.5131	-0.2433	0.2107		16.6464	16.8422	-5.2714
7	-0.0072	-0.1688	0.0800	-0.0693	0.0228		17.0404	-5.3334
8	0.0353	0.8228	-0.3901	0.3378	-0.1111	0.5417		1.6693

[(d'_k - d''_k) (d'_l - d''_l)]									
k/l	1	2	3	4	5	6	7	8	
1	2.8629	-3.2960	-1.5668	-13.8998	-6.0997	6.9034	6.9846	-2.1860	
2		3.7947	1.8038	16.0028	7.0225	-7.9478	-8.0413	2.5168	
3	0.0023		0.8575	7.6071	3.3382	-3.7781	-3.8225	1.1964	
4	0.0537	1.2499		67.4862	29.6151	-33.5172	-33.9115	10.6138	
5	-0.0254	-0.5925	0.2809		12.9960	-14.7084	-14.8814	4.6577	
6	0.0220	0.5131	-0.2433	0.2107		16.6464	16.8422	-5.2714	
7	-0.0072	-0.1688	0.0800	-0.0693	0.0228		17.0404	-5.3334	
8	0.0353	0.8228	-0.3901	0.3378	-0.1111	0.5417		1.6693	

TABLEAU 34

Dispersion $w_{kl} = \sum_{i,j} (x_{ijk} - \bar{x}_{ik})(x_{ijl} - \bar{x}_{il}) / (N - 2)$ et écarts $d_k = \bar{x}_{1k} - \bar{x}_{2k}$
pour la discrimination raciale commune; test t relatif à ces écarts (form. 51)

l/k	1	2	3	4	5	6	7	8
1	13.3207	11.2065	15.3703	9.9445	11.6731	15.9020	4.5187	3.0479
2	..	32.0834	22.3904	12.8297	17.3971	28.4045	9.0770	4.9815
3	..	..	69.9735	27.9695	29.0270	39.1979	13.6228	9.6633
4	..	..	..	46.2672	34.3821	25.8189	10.3477	7.9202
5	..	..	..	..	45.6485	31.0238	11.5371	9.2858
6	..	..	..	..	..	72.9405	24.4136	7.8614
7	..	..	..	..	..	..	17.8151	3.8775
8	..	..	..	..	..	..	..	8.5031
d_k	0.773 °	1.143 °	-0.041 °	4.263 **	1.962 **	-1.801 °	-1.985 **	0.728 *

° différence non significative.
 * différence significative ($P = 0.05$).
 ** différence hautement significative ($P = 0.01$).

TABLEAU 35

Dispersion $[w_{kl}]$ et matrice inverse $[w^{kl}]$ pour la comparaison des fonctions discriminantes raciales réduites aux trois ou quatre caractères principaux

Chiffre supérieur: w_{kl} (dispersion) Chiffre inférieur: w^{kl} (inverse)									
$p = 3$				$p = 4$					
l/k	1	4	7	1	2	4	7	k/l	
1	9.7811 0.1050	2.8948 -0.0092	0.3569 -0.0013	9.7811 0.1182	4.5442 -0.0276	2.8948 -0.0109	0.3569 0.0014	1	
2	— —	— —	— —		19.6790 0.0576	-0.3479 0.0034	1.3396 -0.0057	2	
4		32.5253 0.0319	2.1688 -0.0050			32.5253 0.0321	2.1688 -0.0054	4	
7			13.0515 0.0775				13.0515 0.0781	7	

TABLEAU 36

Matrice de dispersion $[w_{kl}]$ et inverse $[w^{kl}]$ relatives aux comparaisons des individus de type II (♀ et ♂) à ceux de type I (♀ et ♂)

Chiffre supérieur: w_{kl} (dispersion) Chiffre inférieur: w^{kl} (inverse)									
Comparaison WI/WII/RI					Comparaison WI/RII/RI				
l/k	1	2	4	7	1	2	4	7	k/l
1	12.8970 0.1101	9.6271 -0.1284	8.7425 -0.0130	4.6817 -0.0071	13.9186 0.1087	11.2612 -0.0308	10.2724 -0.0141	4.6058 -0.0046	1
2		29.7948 0.0487	11.6212 -0.0036	8.9131 -0.0152		32.2378 0.0472	13.4844 -0.0039	8.9219 -0.0136	2
4			42.9118 0.0303	10.7487 -0.0135			46.4718 0.0281	10.2483 -0.0107	4
7				17.2538 0.0762				17.6135 0.0711	7

TABLEAU 37

Ecart e_k et matrices $[e_k e_l]$ relatives aux comparaisons des individus de type II ($\mathfrak{J}$ et $\mathfrak{Q}$)
à ceux de type I ($\mathfrak{J}$ et $\mathfrak{Q}$)

Comparaison WI/WII/RI					Comparaison WI/RII/RI				
<i>Ecart</i> $e_k = 2\bar{x}_{3k} - \bar{x}_{1k} - \bar{x}_{2k}$									
k	1	2	4	7	1	2	4	7	k
	- 0.799	0.031	1.311	- 1.825	1.509	2.489	- 0.017	0.547	
<i>Matrices</i> $[e_k e_l]$									
l/k	1	2	4	7	1	2	4	7	k/l
1	0.6384	- 0.0248	- 1.0475	1.4582	2.2771	3.7559	- 0.0256	0.8254	1
2	..	0.0010	0.0406	- 0.0566	..	6.1951	- 0.0423	1.3615	2
4	..	..	1.7187	- 2.3926	..	..	0.0003	- 0.0093	4
7	..	..	..	3.3306	..	..	..	0.2992	7

TABLEAU 38

Dispersion $[w_{kl}]$ et *matrice inverse* $[w^{kl}]$ pour le classement individuel
de quatre Walsers occidentaux

Chiffre supérieur: w_{kl} (dispersion) Chiffre inférieur: w^{kl} (inverse)				
l/k	1	2	4	7
1	13.3207 0.1154	11.2065 -0.0334	9.9445 -0.0147	4.5187 -0.0037
2		32.0834 0.0480	12.8297 -0.0030	9.0770 -0.0143
4			46.2672 0.0281	10.3477 -0.0111
7				17.8151 0.0708

